

Cloud Computing

Nova Arquitetura da TI

Manoel Veras

- Vincula Arquitetura de Tecnologia da Informação com CLOUD COMPUTING
- Explica o que é CLOUD COMPUTING e mostra os principais desafios a serem vencidos
- Define as arquiteturas de serviços incluindo IaaS, PaaS, SaaS e apresenta os modelos de implantação público, privado e híbrido
- Mostra o estado atual das ofertas dos principais fornecedores, incluindo Amazon AWS e Microsoft AZURE

Prefácio
Dr. Robert Tozer
Diretor, Operational Intelligence





Manoel Veras

Cloud Computing

Nova Arquitetura da TI

Prefácio

Dr. Robert Tozer

Diretor, Operational Intelligence



Copyright © 2012 por Brasport Livros e Multimídia Ltda.

Todos os direitos reservados. Nenhuma parte deste livro poderá ser reproduzida, sob qualquer meio, especialmente em fotocópia (xerox), sem a permissão, por escrito, da Editora.

Editor: Sergio Martins de Oliveira

Diretora: Rosa Maria Oliveira de Queiroz

Gerente de Produção Editorial: Marina dos Anjos Martins de Oliveira

Revisão: Maria Inês Galvão

Editoração Eletrônica: Abreu's System Ltda.

Capa: Paulo Vermelho

Técnica e muita atenção foram empregadas na produção deste livro. Porém, erros de digitação e/ou impressão podem ocorrer. Qualquer dúvida, inclusive de conceito, solicitamos enviar mensagem para brasport@brasport.com.br, para que nossa equipe, juntamente com o autor, possa esclarecer. A Brasport e o(s) autor(es) não assumem qualquer responsabilidade por eventuais danos ou perdas a pessoas ou bens, originados do uso deste livro.

BRASPORT Livros e Multimídia Ltda.

Rua Pardal Mallet, 23 – Tijuca

20270-280 Rio de Janeiro-RJ

Tels. Fax: (21) 2568.1415/2568.1507

e-mails: brasport@brasport.com.br

vendas@brasport.com.br

editorial@brasport.com.br

site:

www.brasport.com.br

Filial

Av. Paulista, 807 – conj. 915

01311-100 – São Paulo-SP

Tel. Fax (11): 3287.1752

e-mail: filialsp@brasport.com.br

Dedico esta obra aos meus irmãos Palmira, João Hélio, Patrícia e Renata.

“A capacidade de competir de um país é uma manifestação de vontade, uma construção obsessiva, a opção de um povo.”
(Michael Porter).

Agradecimentos

Agradeço a todos que me incentivam a prosseguir. O livro CLOUD COMPUTING é um desdobramento dos livros DATACENTER e VIRTUALIZAÇÃO, mas não menos importante. Estruturá-lo e torná-lo prático sem esquecer-se dos fundamentos teóricos do tema foi um grande desafio. O ponto certo entre a teoria atual sobre a arquitetura de CLOUD COMPUTING (Computação de Nuvem ou Computação em Nuvem ou Computação nas Nuvens, em português) e os aspectos práticos foi o que busquei.

Temos uma carência enorme de material estruturado sobre os assuntos aqui abordados e entendo que a formação de mão de obra neste campo é essencial para o desenvolvimento do Brasil. Sou grato a diversos profissionais e acadêmicos da computação brasileira com quem convivi nestes últimos anos e que me ajudaram a tornar este sonho uma realidade.

Gostaria também de agradecer, mais uma vez, ao Sergio Martins e à Rosa Queiroz, da Editora Brasport, pelo apoio dado ao projeto desde o início.

Manoel Veras

Nota do Autor

As sugestões feitas neste livro devem ser tratadas como linhas orientadoras para profissionais que buscam atuar nas áreas de conhecimento envolvidas e necessitam de um referencial sobre CLOUD COMPUTING. As áreas envolvidas diretamente neste livro são: Tecnologia da Informação (TI), CLOUD COMPUTING, DATACENTERS e VIRTUALIZAÇÃO, normalmente abordadas com diferentes enfoques na Ciência da Computação, na Engenharia da Computação, em Sistemas de Informação e até na Administração de Empresas.

Este é um livro com foco em explicar e mostrar o estado atual da arquitetura CLOUD COMPUTING. CLOUD COMPUTING é uma arquitetura para TI que pode ser considerada uma evolução das arquiteturas MAINFRAME e cliente/servidor. Boas obras sobre CLOUD COMPUTING já existem com foco nesta temática, mas este livro aborda CLOUD COMPUTING do ponto de vista da arquitetura e da infraestrutura. A TI neste caso é tratada como um serviço a ser fornecido internamente (nuvem privada), adquirido externamente (nuvem pública) ou mesmo fornecido em um modelo híbrido. O livro assume que você, caro leitor, tem um conhecimento básico de gestão da TI, hardware, software, redes e seus protocolos.

CLOUD COMPUTING só é efetiva quando possui o(s) DATACENTER(S), seu principal componente, provido com recursos de VIRTUALIZAÇÃO. Integrar os conceitos de CLOUD COMPUTING, DATACENTER e VIRTUALIZAÇÃO é parte da essência deste livro e, sem dúvida, um grande desafio.

Diversas publicações sugerem que o mercado de trabalho para profissionais que lidam com os assuntos aqui tratados só tende a crescer. As organizações precisam de profissionais que entendam qual o papel da TI, compreendam os aspectos relevantes sobre a utilização de CLOUD COMPUTING, do DATACENTER e da VIRTUALIZAÇÃO na nova organização. Estes profissionais também precisam ter uma visão clara sobre as arquiteturas e tecnologias envolvidas com CLOUD COMPUTING e estruturar melhor a decisão de utilizar este modelo de arquitetura como opção para a TI. CLOUD COMPUTING é o presente e o futuro.

Prefácio

Com o crescimento do mercado global de Tecnologia da Informação (TI) cresce também nossa dependência dos serviços de TI. CLOUD COMPUTING é uma abordagem para o compartilhamento de recursos que tem profundas consequências na forma como o mundo faz negócios e como interagimos uns com os outros, e o impacto total ainda não sabemos. Vivemos em uma sociedade cada vez mais dependente da tecnologia e, à medida que aumentamos nossa presença na esfera virtual, os nossos dados tornam-se dispersos em meios que têm pouco controle.

A nuvem oferece flexibilidade através um provisionamento de serviços escalável, aproveitando os avanços da conectividade e as tecnologias de VIRTUALIZAÇÃO e muda a forma de fazer negócios. Ela vai transformar a forma de pensar os DATACENTERS e a concorrência entre operadores irá provavelmente forçar uma consolidação do mercado, deixando alguns operadores com um pequeno número de grandes instalações. A necessidade de minimizar o custo total de propriedade (*Total Cost of Ownership – TCO*) e a pressão sobre recursos continuarão a impulsionar a tendência para densidades de carga crescente, manutenção da eficiência energética e a priorização da alta disponibilidade para os operadores. CLOUD COMPUTING não abrange apenas as mudanças tecnológicas; a indústria precisa de pessoal altamente qualificado para oferecer esses serviços.

Os operadores não podem se dar ao luxo de ignorar o impacto da qualificação da sua equipe sobre o negócio: a maioria das falhas na indústria de DATACENTERS está concentrada no erro humano/gestão (The Uptime Institute, Duffey & Saull 2008, “gestão de risco: o elemento humano”). Também é verdade dizer que as maiores perdas de oportunidades de economia de energia estão relacionadas à inconsciência humana. Criar um ambiente eficaz e uma dinâmica de aprendizagem para equipes operacionais de DATACENTERS é essencial, especialmente quando devemos ter o compartilhamento de conhecimento e colaboração entre as disciplinas de TI e instalações. Podemos aprender algo com as metodologias de Paulo Freire, por exemplo, especialista em educação brasileira que usou técnicas revolucionárias para enfrentar o analfabetismo.

Tive a honra de conhecer Manoel Veras na reunião no Dynamics DATACENTER, conferência em São Paulo em 2009. Quando o conheci eu estava fascinado por sua abordagem à educação, promoção de um ambiente verdadeiramente inspirador de aprendizagem em organizações de DATACENTERS. Esta é a chave para a redução do custo total de propriedade de um DATACENTER em termos de despesas de capital, custos operacionais e confiabilidade.

Este livro fornece insights sobre as questões e desafios que existem para a nossa indústria a nível internacional e nos ajuda a entender as mudanças e como terão de se adaptar. É dividido em três seções que abordam aspectos do negócio, a infraestrutura de nuvem (ou seja, o DATACENTER) e serviços em nuvem (IaaS, PaaS, SaaS). Tenho certeza de que você vai encontrar informações instigantes sobre todos estes assuntos.

Dr. Robert Tozer
Diretor, Operational Intelligence

Sumário

Introdução

Objetivos

Estrutura

PARTE I: ASPECTOS GERAIS

1. Tecnologia da Informação e Cloud Computing

1.1. Introdução

1.2. Financiamento da TI

1.3. Alinhamento Estratégico

1.4. Arquitetura Empresarial

1.5. Arquitetura de TI

1.6. Conceito na Prática: Oracle Enterprise Architecture Framework

1.7. TI como Serviço (TlaaS) e CLOUD COMPUTING

1.8. Referências Bibliográficas

2. Visão Geral

2.1. Introdução

2.2. Conceito

2.3. Características Essenciais

2.4. Modelos de Serviço

2.5. Modelos de Implantação

2.6. Conceito na Prática: Modelo para Segurança da CSA

2.7. Arquitetura Multitenancy ou Multi-Inquilino

2.8. Iniciativas

2.9. Referências Bibliográficas

3. Benefícios e Riscos

3.1. Introdução

3.2. O Benefício da Economia de Escala

3.2.1. Economia de Escala do Lado do Fornecimento

3.2.2. Economia de Escala do Lado da Demanda

3.2.3. Economia de Escala da Arquitetura Multitenancy

3.3. Outros Benefícios

- 3.4. Riscos
- 3.5. Consolidação dos Riscos
- 3.6. Referências Bibliográficas

4. Tomada de Decisão

- 4.1. Introdução
- 4.2. Governança da TI
- 4.3. Governança de TI e CLOUD COMPUTING
- 4.4. Terceirização da TI e CLOUD COMPUTING
 - 4.4.1. Introdução
 - 4.4.2. Classificação da Terceirização de TI
 - 4.4.3. Benefícios da Terceirização da TI
 - 4.4.4. Riscos da Terceirização de TI
- 4.5. Governança de TI e Arquitetura Empresarial
- 4.6. Terceirização da TI e Arquitetura Empresarial
- 4.7. Conceito na Prática: CLOUD COMPUTING no Governo Americano
- 4.8. Seleção do Provedor de CLOUD COMPUTING
- 4.9. Referências Bibliográficas

PARTE II: INFRAESTRUTURA DE NUVEM

5. DATACENTER: Aspectos Gerais

- 5.1. Introdução
- 5.2. UPTIME Institute
 - 5.2.1. Introdução
 - 5.2.2. Norma TIA-942
 - 5.2.3. Camadas (TIERS)
 - 5.2.4. Certificações TIER
- 5.3. Custo do DATACENTER
- 5.4. Padronização do DATACENTER
 - 5.4.1. CHASSI
 - 5.4.2. RACK
 - 5.4.3. CONTAINER
- 5.5. Instalação e Construção do DATACENTER
 - 5.5.1. Introdução
 - 5.5.2. Modularidade
- 5.6. Visão Geral dos DATACENTERS em CONTAINERS
 - 5.6.1. Introdução
 - 5.6.2. Seleção de CONTAINERS
 - 5.6.3 DATACENTER em CONTAINER (CDC) versus DATACENTER Tradicional (TDC)
- 5.7. Segurança Física do DATACENTER
- 5.8. Gerenciamento do DATACENTER
 - 5.8.1. DCIM
- 5.9. Referências Bibliográficas

6. DATACENTER: Eficiência Energética

6.1. Introdução

6.2. Eficiência Energética do DATACENTER

6.3. Equação Energética do DATACENTER

6.4. O Green Grid

6.4.1. Introdução

6.4.2. PUE e DCiE

6.4.3. PUE 2

6.4.4. Novos Indicadores CUE e WUE

6.4.5. Modelo de Maturidade do DATACENTER (DCMM)

6.5. Conceito na Prática: Open Compute Project

6.6. Conceito na Prática: Google DATACENTERS

6.7. Conceito na Prática: HP POD 240a

6.8. Referências Bibliográficas

7. DATACENTER: Arquitetura

7.1. Introdução

7.2. Arquitetura do DATACENTER

7.3. Arquitetura Virtual do DATACENTER

7.4. VIRTUALIZAÇÃO

7.4.1. Conceito

7.4.2. Efeitos

7.4.3. Conceito na Prática: VMware vFabric CLOUD Application Platform

7.5. CLUSTERIZAÇÃO

7.5.1. Introdução

7.5.2. Clusters de Balanceamento de Carga e de Alta Disponibilidade

7.5.3. Cluster de Alta Performance

7.5.4. Top 500 Supercomputers Site

7.5.5. Clusters em Grid

7.5.6. Conceito na Prática: Oracle Exadata

7.6. Blocos de Construção da TI

7.6.1. Introdução

7.6.2. Conceito na Prática: VCE da VMware, Cisco e EMC

7.6.3. Padrão FCoE

7.6.4. Conceito na Prática: Open FCoE da Intel

7.7. Referências Bibliográficas

PARTE III: SERVIÇOS DE NUVEM

8. Infraestrutura como Serviço (IaaS)

8.1. Introdução

8.2. Conceito na Prática: Amazon AWS

8.2.1. Introdução

8.2.2. Serviços AWS

- 8.2.3. Centros de Suporte dos Serviços AWS
- 8.2.4. Funcionamento do Amazon AWS EC2
- 8.2.5. Aplicações Tolerantes a Falhas no AWS
- 8.2.6. Segurança no AWS
- 8.2.7. Capacidade do AWS
- 8.2.8. Precificação do AWS

8.3. Referências Bibliográficas

9. Plataforma como Serviço (PaaS)

9.1. Introdução

9.2. Conceito na Prática: Windows Azure

- 9.2.1. Introdução
- 9.2.2. Windows Azure
- 9.2.3. SQL Azure
- 9.2.4. Windows Azure AppFabric
- 9.2.5. Windows Azure Marketplace
- 9.2.6. Precificação do Windows Azure
- 9.2.7. Segurança do Windows Azure
- 9.2.8. Capacidade do Windows Azure
- 9.2.9. Windows Azure versus Amazon AWS

9.3. Referências Bibliográficas

10. Software como Serviço (SaaS)

10.1. Introdução

10.2. Benefícios do SaaS

- 10.2.1. Melhor Gerenciamento dos Riscos da Aquisição de Software
- 10.2.2. Mudança no Foco da TI

10.3. Diferenças entre Software Convencional e SaaS

10.4. Considerações para Adotar o SaaS

10.5. Abordagens para a Arquitetura Multitenancy

- 10.5.1. Introdução
- 10.5.2. Banco de Dados Separado
- 10.5.3. Banco de Dados Compartilhado, Esquemas Separados
- 10.5.4. Banco de Dados Compartilhado, Esquemas Compartilhados
- 10.5.5. Considerações Econômicas
- 10.5.6. Considerações de Segurança
- 10.5.7. Considerações sobre Tenants
- 10.5.8. Considerações sobre Mudanças
- 10.5.9. Considerações sobre Habilidades Necessárias
- 10.5.10. Qualidades de uma Aplicação SaaS

10.6. Conceito na Prática: Force.com

- 10.6.1. Introdução
- 10.6.2. Arquitetura Metadata-Driven

10.7. Conceito na Prática: Office 365 da Microsoft

10.7.1. SharePoint Online

10.7.2. Exchange Online

10.7.3. Gerenciamento e Migração do BPOS para Office 365

10.8. Referências Bibliográficas

Introdução

CLOUD COMPUTING trata de mudança. Mudança que está remodelando o setor de TI, segundo Nicholas Carr¹. A ideia central colocada por NIC é que a TI vai ser fornecida como serviço público logo mais adiante, como aconteceu com a energia. Esta nova forma de entregar e receber a TI é a que se convencionou chamar de CLOUD COMPUTING.

A VIRTUALIZAÇÃO ajudou as empresas a usar os recursos de hardware com mais eficiência. Ela possibilitou desacoplar o ambiente do software do hardware. Agora, os servidores existem como se fossem um único arquivo, uma máquina virtual. É possível movê-los de um hardware para o outro, duplicá-los quando desejar e criar uma infraestrutura mais escalonável e flexível.

Os DATACENTERS aproveitaram a VIRTUALIZAÇÃO e tornaram-se mais disponíveis e mais eficientes. Os recursos agora são mais bem utilizados e as capacidades da TI mais bem aproveitadas.

CLOUD COMPUTING aumentou ainda mais esse nível de eficiência e agilidade atingido pela VIRTUALIZAÇÃO dos DATACENTERS. Por meio de recursos em pool, diversidade geográfica e conectividade universal, CLOUD COMPUTING facilita o fornecimento de softwares hospedados, plataformas e da infraestrutura como um serviço. Ela é, ao mesmo tempo, uma nova plataforma tecnológica e uma nova arquitetura de TI.

CLOUD COMPUTING já é uma realidade. Diversas formas de uso e novas aplicações surgem e a demanda por profissionais que entendam a mudança e preparem as organizações para este novo paradigma da computação só aumenta.

Este livro trata da arquitetura de CLOUD COMPUTING. São trazidos aqui aspectos e conceitos importantes que contribuem para a formação de profissionais na área de Tecnologia da Informação (TI) com foco nesta nova arquitetura. Venho estudando o assunto há cinco anos e só agora senti que poderia produzir um texto útil, com conteúdo, que pudesse servir de referência para profissionais e estudantes da área.

Qual a linha de base estabelecida para o livro? Partiu-se do genérico, associando a arquitetura empresarial à arquitetura de CLOUD COMPUTING, indo até o específico, tratando de questões puramente técnicas relacionadas à arquitetura de TI e tecnologias envolvidas. Uma dificuldade natural de um livro com este foco é conseguir sequenciar os assuntos de forma a fazer com que o leitor avance passo a passo. Procurou-se construir os assuntos na melhor sequência possível, mas eventualmente é preciso chamar um conceito que só será explicado posteriormente. Este aspecto deve ser considerado durante a leitura.

Objetivos

Os principais objetivos deste livro são:

- Auxiliar no crescimento da área de TI no Brasil.
- Ajudar a formar mão de obra qualificada em TI no Brasil.
- Auxiliar consultores de TI no exercício da profissão.

Estrutura

O livro possui dez capítulos. A ideia é que os assuntos tratados nos capítulos tenham certa independência, mesmo que fazendo parte de uma sequência lógica e que assim permitam que o leitor possa ler um único capítulo.

Vale salientar que o aspecto prático é sempre considerado e o livro traz diversos exemplos de casos e dicas de implementações reais das tecnologias citadas.

Importante deixar claro que as seções **Conceito na Prática** são baseadas em informações fornecidas pelos fabricantes em artigos públicos, folhas de especificação (*spec sheets*) ou em seus sites, e não são originadas pelo autor.

Neste livro, aplicação e aplicativo são utilizados como sinônimos.

Optou-se também por utilizar o termo CLOUD COMPUTING e não computação de nuvem ou computação em nuvem ou computação nas nuvens ao longo do livro.

As partes do livro são divididas em:

- Aspectos Gerais: Capítulos 1, 2, 3 e 4.
- Infraestrutura de Nuvem: Capítulos 5, 6 e 7.
- Serviços de Nuvem: Capítulos 8, 9 e 10.

A descrição dos capítulos e as respectivas partes são mostradas na **Figura 0-1**.

1

Capítulo 1
Tecnologia da
Informação e
CLOUD
COMPUTING

Capítulo 2
Aspectos
Gerais

Capítulo 3
Benefícios
e
Riscos

Capítulo 4
Tomada de
Decisão

2

Capítulo 5
DATACENTER:
Aspectos Gerais

Capítulo 6
DATACENTER
Eficiência
Energética

Capítulo 7
DATACENTER
Arquitetura

3

Capítulo 8
Infraestrutura
como Serviço -
IaaS

Capítulo 9
Plataforma
como Serviço -
PaaS

Capítulo 10
Software como
Serviço - SaaS

Figura 0-1 – Capítulos do Livro

PARTE I:

ASPECTOS GERAIS

1. Tecnologia da Informação e Cloud Computing

1.1. Introdução

A Tecnologia da Informação (TI) é a tecnologia que suporta a informação, seu processamento e armazenamento, utilizada para objetivos diversos. Acredita-se que a TI é fundamental para a melhoria da competitividade de uma organização.

Com o avanço do uso de processos empresariais que utilizam a TI em grande escala, ela tornou-se a “espinha dorsal” para muitos negócios e o próprio negócio de outros negócios.

No Brasil, por exemplo, de acordo com a consultoria IDC, a indústria de Tecnologia da Informação emprega 600 mil pessoas e movimentará o equivalente a US\$ 39 bilhões em hardware, software e serviços no ano de 2011. Computada a TI utilizada pelo governo e em outras atividades da economia, o setor tem um peso relativo de 3,2% do PIB, com um mercado total de cerca de US\$ 68 bilhões já em 2011.

Importante ressaltar que a TI se encontra em diferentes estágios em diferentes organizações. Sua maior ou menor importância vai depender de como ela é utilizada e da maturidade deste uso. Em certas organizações a TI é vista e tratada como custo, em outras a TI é vista como estratégica e geradora de valor.

Considerando que a TI é importante, mesmo que muitas vezes vista como um custo, a grande questão hoje é saber como fazer a TI contribuir para a melhoria do desempenho empresarial, considerando as suas diferentes formas de uso na organização e o fato dela, a TI, permear todos os setores da organização, dificultando a própria monitoração do seu uso e do seu valor agregado.

As organizações são coletâneas de grandes processos. Processos que devem responder a demandas das mais diversas. Fusões e aquisições de empresas, por exemplo, são movimentos cada vez mais comuns que alteram estratégias preestabelecidas. Estas alterações reconfiguram processos que devem refletir a nova organização.

Se processos de negócio são alterados em função de estratégias que se modificam constantemente devido principalmente às mudanças do ambiente, que tal tratar de ter uma TI flexível que permita a rápida reconfiguração da organização? Será que a forma que a TI existe atualmente em boa parte das organizações permite obter a flexibilidade necessária? Será que a forma que a TI está configurada dentro de boa parte das organizações permite a entrega de serviços, resultante da qualidade dos seus processos internos, com a qualidade negociada? Será que a forma que a TI é financiada é a mais adequada?

Mas como tornar a TI flexível? Organizações, em sua grande maioria, possuem um legado, um conjunto de aplicativos que se comunicam de forma precária e dados duplicados. Romper com este passado é um ato de inteligência, mas na maioria dos casos não é uma tarefa trivial, pois a organização está em pleno funcionamento e qualquer migração de sistemas ou mesmo atualização pode ser motivo para haver perda de dados e *downtime* dos aplicativos. Também pode haver falta de

recursos para novos projetos. A infraestrutura, por sua vez, precisa ser repensada, pois com aplicativos construídos para serem acessados por usuários que estão em qualquer lugar do mundo, a infraestrutura baseada em um acesso quase que exclusivamente local não serve mais.

Organizações assim, em sua grande maioria, quase sempre focam na operação do dia a dia e esquecem a inovação. Os aplicativos e a infraestrutura consomem boa parte do tempo dos funcionários da TI e também consomem quase todos os recursos alocados para a TI. O diretor de TI, por sua vez, só trata de questões puramente operacionais.

Como então repensar a TI?

A TI tem quatro grandes partes: os sistemas de informação (conjunto de aplicativos), a arquitetura, a infraestrutura e a gestão. Considera-se aqui que as pessoas que suportam a TI fazem parte da infraestrutura e que a governança é parte da gestão.

A arquitetura de TI, explicada detalhadamente mais à frente neste capítulo, é normalmente formada por duas grandes partes: a arquitetura dos aplicativos e a arquitetura da infraestrutura.

A arquitetura dos aplicativos trata do desenho dos aplicativos, da forma de construção e do seu reaproveitamento. A ideia hoje é que componentes de software que fazem parte do aplicativo possam ser reaproveitados em novos desenvolvimentos, aumentando a eficiência da TI.

Os aplicativos dão vida aos processos de negócio e boa parte das informações que fazem parte dos processos é gerada e tratada por estes aplicativos.

A arquitetura da infraestrutura trata do desenho da infraestrutura. As partes de infraestrutura precisam ser pensadas, de forma a permitir o ganho de escala e a otimização de recursos. Parte deste esforço passa pela modularidade das soluções de infraestrutura, que permitem, por sua vez, obter a flexibilidade necessária.

A infraestrutura é o alicerce para os aplicativos e sustenta o modelo operacional, modelo que define como os processos estão integrados e padronizados.

A infraestrutura de TI, como qualquer outra infraestrutura, tem o papel de possibilitar que a organização funcione e cresça sem grandes interrupções. As organizações dependem cada vez mais da infraestrutura de TI, na medida em que trocam processos de negócios analógicos por processos digitais que são a base do seu modelo operacional.

Vale ressaltar que a infraestrutura de TI de hoje é mais complexa do que a infraestrutura de TI de alguns anos atrás, pois é uma combinação de infraestrutura privada (redes e dispositivos que conectam unidade de negócio, organização, setor de atuação) e pública (normalmente a Internet). A Internet é uma via pública, e a garantia de serviços nesta rede é uma tarefa complexa. As opções referentes à infraestrutura de TI são muitas e as decisões precisam ser criteriosas, pois envolvem altos investimentos.

A execução da estratégia empresarial, ancorada no modelo operacional, acaba dependendo da condição que a infraestrutura e os aplicativos proporcionam. Alguns autores reforçam que a infraestrutura de TI, no final das contas, é quem também responde pela condição de inovar de uma organização nos dias atuais, mesmo que no nível operacional².

A grande questão é modificar a TI, sua gestão, infraestrutura e arquitetura para que ela, a TI, suporte de forma flexível os processos de negócio e por sua vez a estratégia.

A **Figura 1-1** ilustra a relação entre os componentes da TI os processos empresariais e o desempenho empresarial. Todos os recursos ilustrados devem estar alinhados para melhorar o desempenho empresarial.

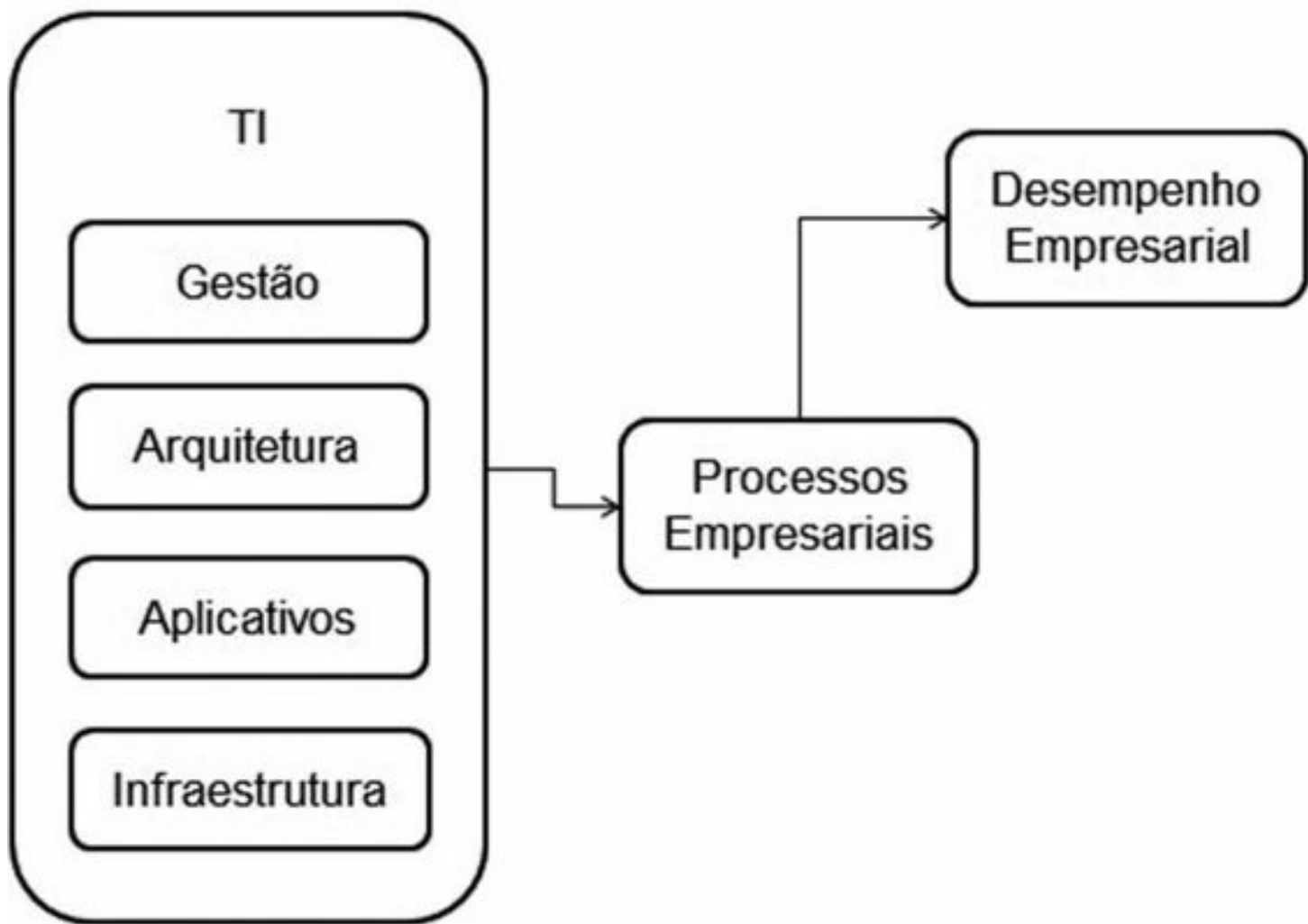


Figura 1-1 – TI e Desempenho Empresarial

Outro conceito fundamental é o de governança de TI, tratado aqui como parte da gestão da TI. A governança de TI deve alocar a responsabilidade pela definição, provisionamento e precificação dos serviços compartilhados de TI, que decorrem da infraestrutura, buscando alinhar o nível destes serviços com as recomendações definidas na estratégia de TI para as aplicações. A governança de TI decorre da estratégia e da gestão da TI, que, por sua vez, devem estar de acordo com a estratégia da organização.

A estrutura de governança também deve ser repensada em boa parte das organizações para considerar assim o papel estratégico da informação e da tecnologia que a suporta.

Com tudo isto, aplicativos, arquitetura, infraestrutura e gestão inadequadas, a TI ainda precisa cuidar de novos projetos. Para complicar, a dependência da TI só aumentará. O IDC estima que, em 2020, o universo digital (toda informação criada e replicada em formato digital) será 44 vezes maior que em 2009, saindo de 0,8 ZB (1 ZB=1.000.000.000.000 GB) para 35 ZB. Pode-se assim ter uma ideia de como as organizações vão depender cada vez mais da infraestrutura e dos aplicativos para operar. Cerca de 25% deste universo é de informação empresarial. A **Figura 1-2** ilustra o provável

crescimento da base digital de informações, segundo este documento do IDC.

O universo digital atual também é marcado pelo “BIG DATA”. O que seria o BIG DATA? Recentemente Tom White³ cunhou o termo BIG DATA para *datasets* cujo tamanho estão fora do controle dos softwares de gerenciamento de banco de dados. Softwares de gerenciamento de banco de dados capturam, armazenam, gerenciam e analisam dados. BIG DATA não tem um tamanho específico, pois se considera que os dados continuarão crescendo e, mesmo que os softwares consigam gerenciá-los em certo momento, logo depois não mais o farão.

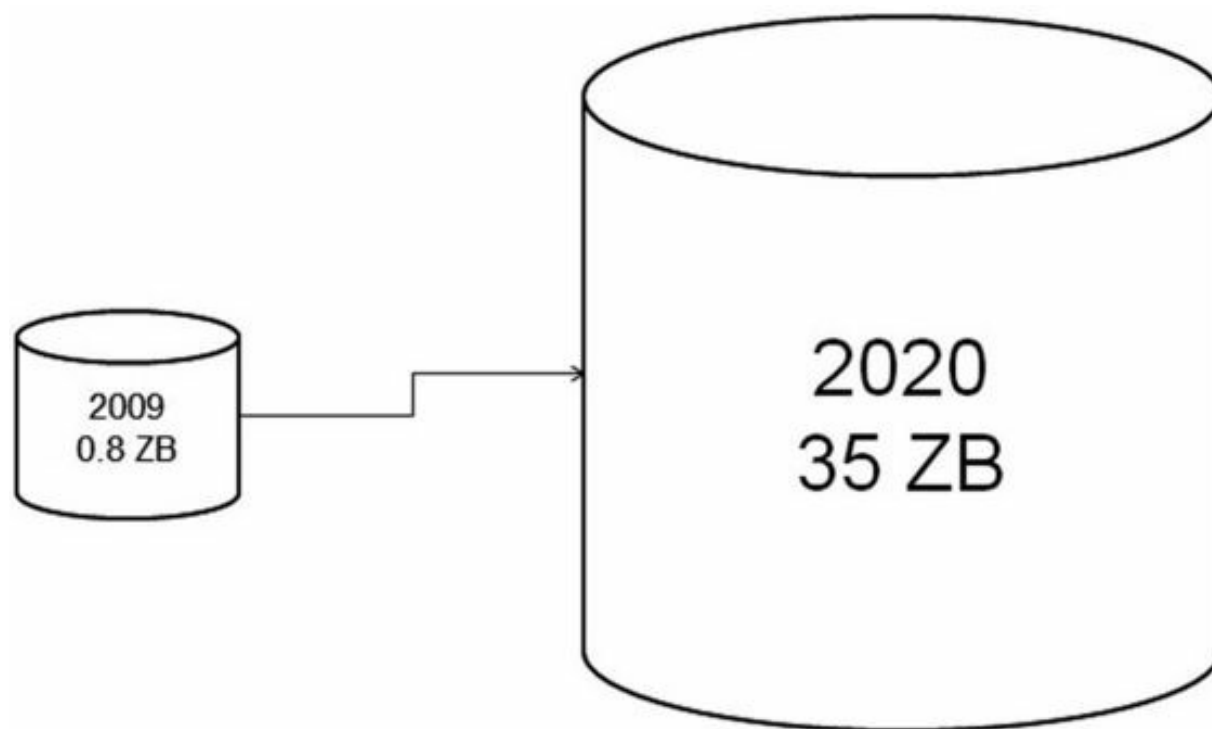


Figura 1-2 – Expansão do Universo Digital

Outro aspecto relevante que torna a informação digital abundante é a disponibilização de banda larga. A introdução da banda larga em grande escala em vários países, incluindo o Brasil, reforça também a importância da TI como alicerce importante do mundo baseado em informação. O acesso em banda larga é caracterizado pela disponibilização de infraestrutura de TI que possibilita tráfego de informações contínuo, ininterrupto e com capacidade suficiente para as aplicações de dados, voz e vídeo. Os Estados Unidos, por exemplo, definiu como marco o valor de 100 Mbps como velocidade de conexão de *download* para cem milhões de residências americanas até 2020. O Brasil também já possui seu plano nacional de banda larga. O avanço da adoção da banda larga sinaliza a opção digital do mundo contemporâneo e reforça a necessidade de governos e organizações privadas planejarem a utilização de uma “plataforma digital” como fator de competitividade nacional.

A dependência da organização da infraestrutura e dos aplicativos de negócio exige cada vez mais a participação dos diretores e gestores de TI em questões de planejamento e decisões de investimento. Esta participação normalmente encontra uma barreira em boa parte das organizações, pois normalmente é mais fácil para um executivo de alto escalão entender um investimento em marketing do que entender o investimento em TI. De qualquer forma, aos poucos, o gestor de TI vem

aumentando o seu espaço dentro das organizações e suas atribuições vêm mudando⁴.

O modelo de governança de TI adotado pela maioria das grandes organizações ilustra o fato de que a TI talvez não tenha ainda assumido o seu verdadeiro papel nas organizações e ilustra também o fato do CFO (*Chief Financial Officer*), por questões de controle e de responsabilidade, estar voltando a liderar as decisões de investimento em TI. O CFO autoriza 26% de todos os investimentos em TI e 51% quando combinado com o CIO, segundo revela a edição 2011 do Financial Executives International (FEI) Technology Study, que mostra também um aumento significativo da quantidade de CIOs que passaram a se reportar aos CFOs, em relação à edição de 2010. O estudo ouviu executivos (75% deles CFOs) de 344 empresas de diversos segmentos econômicos, 49% delas com operações globais. Em 46% das organizações a área de TI se reporta diretamente ao CFO (em 2010 eram 42%). E em 45% delas, o CFO lidera a estratégia de investimento em tecnologia, por ser o único decisor (7%) ou por liderar a equipe que toma decisões sobre tecnologia e TIs (38%). Só em 5% das empresas o CIO continua soberano em relação às decisões de investimentos em tecnologia. Portanto, existe aqui um dilema. Tecnologia da Informação parece ser fundamental. Por outro lado, o posicionamento usual do diretor de TI (CIO) não reflete a importância da TI. O repensar da TI passa principalmente pela mudança do seu atual modelo de financiamento.

1.2. Financiamento da TI

Pesquisas indicam que o orçamento de TI para grandes organizações nos Estados Unidos tem se mantido constante ou mesmo reduzido nos últimos cinco anos, mesmo com demandas crescentes de serviços de provimento de informação. Isto vale também, com algumas exceções, para empresas em outros países. O congelamento dos orçamentos para a TI ocorreu devido a diversos fatores, incluindo os poucos resultados obtidos ou mesmo comunicados pela TI.

A manutenção do orçamento sinaliza que é preciso melhorar ainda mais a eficiência da TI e assim gastar menos com a operação da TI para que sobre dinheiro para novos projetos relacionados à inovação. Pesquisas apontam um custo operacional que, em média, consome 80% dos recursos contra 20% que são utilizados para novos projetos. A ideia é alterar esta relação, aumentando a parte que cabe a novos projetos e, portanto, a parte que cabe à inovação.

Sim, e aí? O que fazer? É preciso transformar a TI. Segundo Weill e Ross (2010), abordagens utilizadas para mudar a TI nos últimos anos se mostraram inadequadas. Entre elas destacaram-se:

- Pôr mais dinheiro nos problemas de TI: em muitos casos, esta opção só aumentou os gastos e não os benefícios.
- Cortar drasticamente os gastos com TI: no curto prazo é uma saída que força o diálogo sobre as prioridades da empresa, mas pode minar a competitividade no longo prazo.
- Demitir o CIO: se for só para achar culpado, não resolve. O CIO, muitas vezes, não teve suas responsabilidades aceitas pelo restante da equipe administrativa e, portanto, não conseguiu exercer o seu papel.
- Terceirizar o problema da TI: pode não ser a solução, se não houver mudança dos hábitos em relação à TI. Os custos e serviços possivelmente não melhorarão significativamente

se as pessoas de negócio não modificarem os hábitos em relação à TI.

- Remover sistemas legados e substituí-los por um grande sistema integrado desenvolvido externamente (ERP ou coisa parecida): o sistema integrado resolve parte do problema, mas se não houver mudança na forma da gestão o sistema por si só não mudará o rumo da empresa.

A transformação, segundo Weill e Ross, começa por mudar o modelo de financiamento da TI. A sugestão dos autores citados é construir uma organização com conhecimento em TI através da criação de uma plataforma digitalizada (equivalente à arquitetura empresarial, termo explicado posteriormente). TI neste novo modelo seria a base da capacidade competitiva da organização, um verdadeiro ativo estratégico. Nas cinco abordagens citadas que tentam resolver o problema da TI, TI é um passivo estratégico.

Weill e Ross (2010) citam três componentes importantes para um novo modelo de financiamento da TI:

- Altos executivos devem estabelecer prioridades e critérios claros para os investimentos em TI.
- A gerência deve desenvolver um processo transparente para avaliar os projetos potenciais.
- Devem ser alocados recursos e seus impactos monitorados. A organização deve utilizar o aprendizado para direcionar investimentos futuros.

O modelo de financiamento da TI será também muito impactado pela adoção da CLOUD COMPUTING, parte da nova forma de pensar a TI, conforme será visto adiante.

1.3. Alinhamento Estratégico

Um dos conceitos-chaves da transformação da TI é o conceito de alinhamento estratégico, que é um componente central da governança de TI e permite, quando bem feito, executar os projetos que são priorizados de acordo com a estratégia. Com o alinhamento, a ideia é deixar de lado o tradicional método de tentativa e erro, muito comum nos ambientes que não utilizam o planejamento como ferramenta de gestão. Mas realizar na prática o alinhamento estratégico não é uma tarefa trivial.

O alinhamento estratégico foca em garantir a ligação entre os planos de negócio e de TI, definindo, mantendo e validando a proposta de valor de TI, alinhando as operações de TI com as operações da organização.

Em muitos casos, mesmo com a execução do Plano Estratégico de TI em conformidade com o Planejamento Estratégico da Organização, o dito alinhamento, a organização continua tendo problemas com a operação e o dia a dia.

Em TI é comum existirem três planos que se completam: o plano estratégico de TI (PETI), que define os objetivos e projetos estratégicos, os planos táticos de TI (PTTI), que tratam dos planos de execução dos projetos prioritários e da alocação de recursos, e o Plano Diretor de TI (PDTI), gerado após o Plano Estratégico de TI e da definição dos Planos Táticos de TI (PTTI), que seria

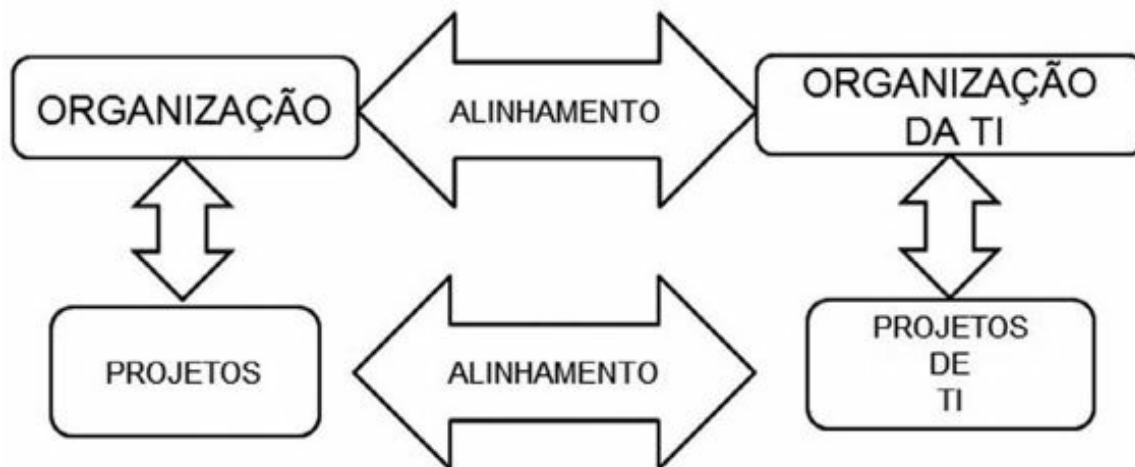


Figura 1-4 – Alinhamento Estratégico

O conceito de portfólio de projetos de TI envolve a soma total dos investimentos em TI, incluindo contratação de serviços e aquisições de hardware, software, redes e contratação de pessoal. O portfólio de projetos de TI pode ser gerenciado como um portfólio financeiro, pesando-se riscos e benefícios dos projetos para o atingimento das metas empresariais.

A abordagem de portfólio de investimentos em TI define quatro classes de investimentos para TI:

- **Infraestrutura de TI:** trata de prover serviços compartilhados e integração;
- **Aplicações Estratégicas:** tratam de prover vantagem competitiva;
- **Aplicações Informacionais:** tratam de prover informações analíticas;
- **Aplicações Transacionais:** tratam de processar transações e cortar custos.

A grande questão é a definição de quanto dos investimentos para a TI vão para as três classes de aplicações e para infraestrutura, incluindo o staff. A abordagem de portfólio permite distribuir o investimento baseado na priorização imposta pela estratégia da organização.

O problema desta abordagem é que os projetos prioritários de hoje não são os projetos prioritários daqui a seis meses. O ambiente se encarrega de complicar este aspecto. E outra: alguns projetos importantes não conseguem ser previstos, simplesmente surgem e precisam ser executados.

Com a adoção da CLOUD COMPUTING, parte do repensar da TI, boa parte das aquisições de TI necessárias à execução de novos projetos será substituída por contratação de serviços pagos pelo uso, conforme será visto adiante.

1.4. Arquitetura Empresarial

Como estão as prioridades dos executivos de TI? Qual o papel da TI nas grandes organizações? 2.014 CIOs (*Chief Information Officer*) foram entrevistados pelo Gartner em 2010. Eles comandam um orçamento de US\$ 160 bilhões e atuam em 38 indústrias de 50 países. Segundo estes CIOs, as prioridades para 2011 são:

- Crescimento do negócio.
- Atrair e reter clientes.
- Reduzir custos.
- Criar novos produtos e serviços (inovação).

As prioridades são estas, mas o fato de serem prioridades não garante que a TI vai conseguir atingir estes quatro objetivos. Isto acontece, muitas vezes, devido à TI ainda ser reativa em boa parte das organizações, conforme anteriormente citado. A operação da TI consome boa parte do orçamento e do tempo do pessoal de TI. Novas demandas em boa parte das vezes geram novos projetos e novas orientações tecnológicas, o que pode dificultar a obtenção do alinhamento.

A ideia de autores como Ross, Weill e Robertson (2006) é que a organização deve construir uma arquitetura empresarial (arquitetura corporativa ou plataforma digitalizada) que reflita os requisitos de padronização e integração dos processos de negócio, o que convencionam chamar de modelo operacional. Esta arquitetura é que fornecerá a agilidade necessária aos negócios e a execução da estratégia. Com esta visão, o planejamento estratégico deixa de ser prioritário e parte-se para criar uma arquitetura estratégica como prioridade.

Os autores citados reforçam que a forma direta de fazer o alinhamento entre os objetivos do negócio e as capacidades da TI tem falhado. As razões seriam que nem sempre a estratégia do negócio é clara para determinar a ação, a empresa cria soluções de TI, e não capacidades da TI. Os processos de implementação são fragmentados e sequenciais; acaba que cada solução de TI criada para responder a uma demanda organizacional é baseada em uma tecnologia diferente. Também a TI quase sempre é o gargalo organizacional, pois está sempre reagindo à demanda organizacional.

Além disso, a velocidade das mudanças ditadas por aquisições e fusões de empresas que possuem arquiteturas de TI diferentes torna este processo ainda mais complicado.

A arquitetura de TI de muitas organizações é uma confusão de aplicações, dados, infraestruturas e tecnologias distintas. A **Figura 1-5** ilustra a abordagem tradicional para soluções de TI.

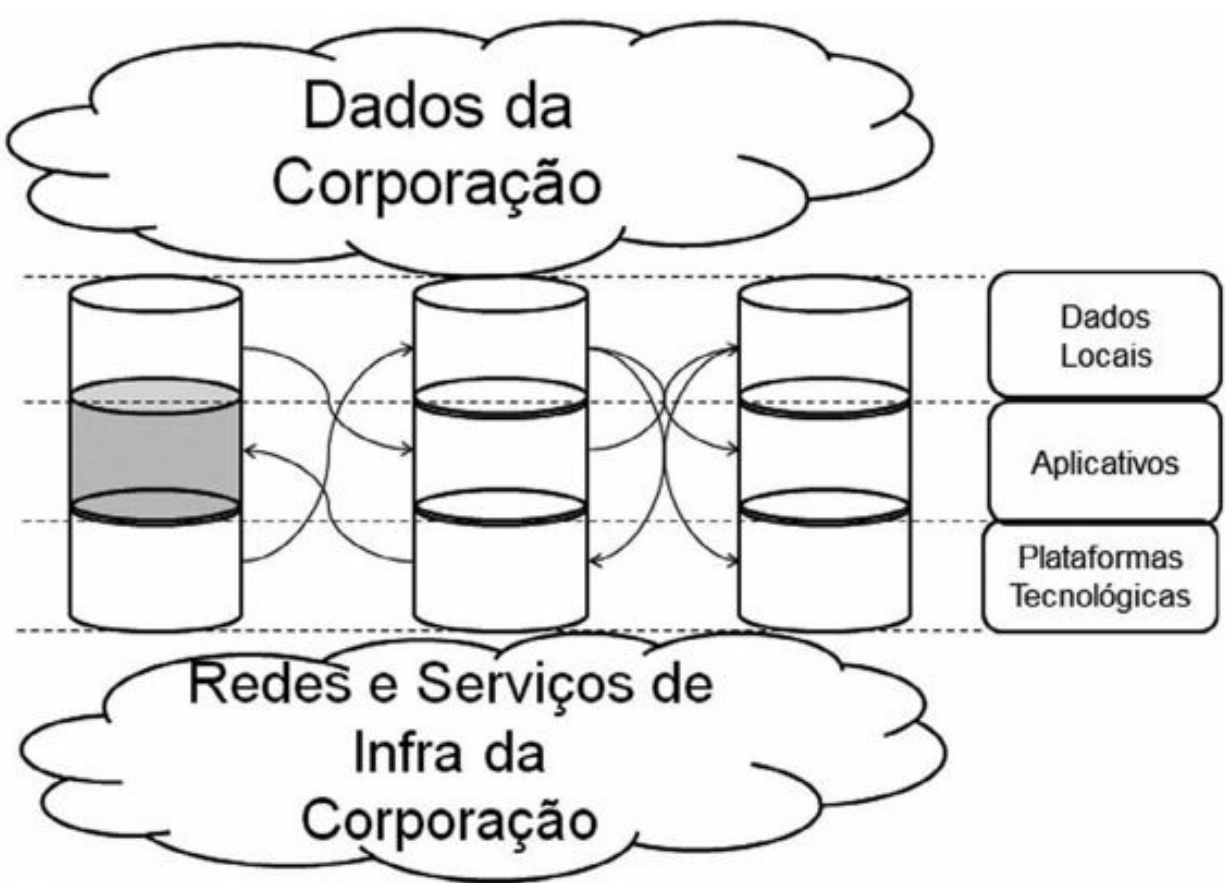


Figura 1-5 – Arquitetura Tradicional da TI.

As soluções encontradas para resolver estas questões já se mostraram ruins e foram citadas anteriormente. O conceito de arquitetura empresarial (Enterprise Architecture – EA) prioriza a arquitetura, e não o planejamento. A ideia central é que a empresa adote um alicerce de execução que resulte da seleção cuidadosa de processos e TI com adequado nível de padronização e integração. A empresa precisaria ter um modelo operacional, e a arquitetura empresarial seria o reflexo deste modelo. Na prática, o modelo operacional seria implementado por meio da arquitetura empresarial. Além disso, mecanismos de governança deveriam assegurar que os projetos de negócio e de TI atinjam os objetivos. Neste caso, os projetos seriam tratados um a um com o apoio da arquitetura.

A ideia de utilizar a arquitetura empresarial como base para a execução da estratégia difere do modelo sugerido anteriormente, que prioriza o planejamento estratégico, conforme ilustra a **Figura 1-6**.

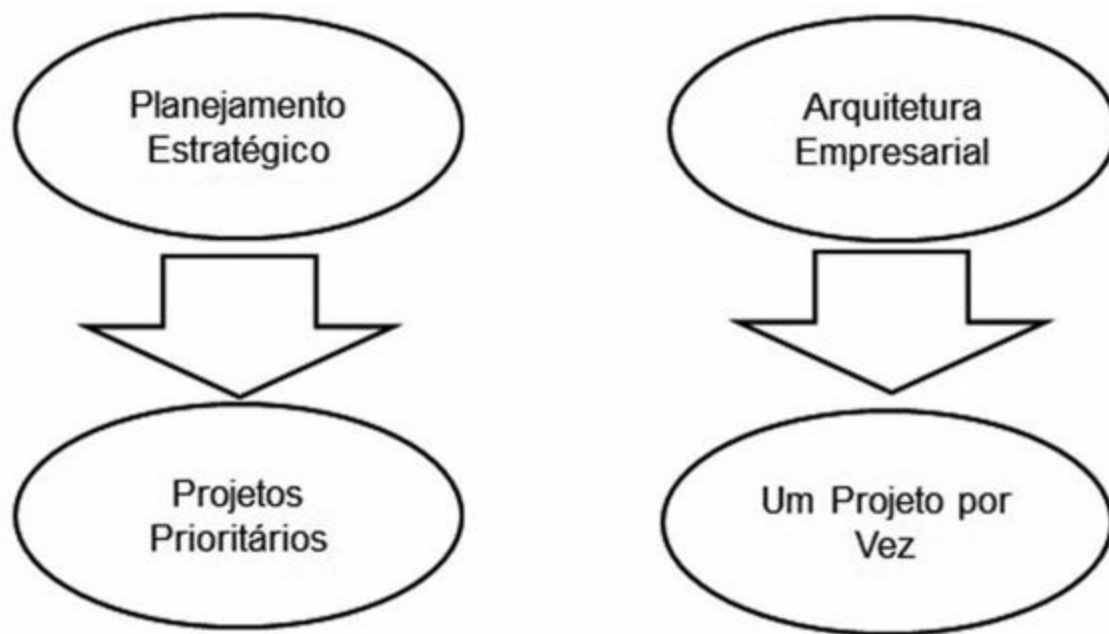


Figura 1-6 – Planejamento Estratégico ou Arquitetura Empresarial

O modelo operacional sugerido por Ross, Weill e Robertson (2006) é de responsabilidade da alta gerência, pois define o lucro e o crescimento a serem alcançados pelo negócio. Os autores sugerem quatro modelos operacionais básicos para facilitar o entendimento do que seja a arquitetura empresarial. Estes modelos são relacionados na **Figura 1-7**.

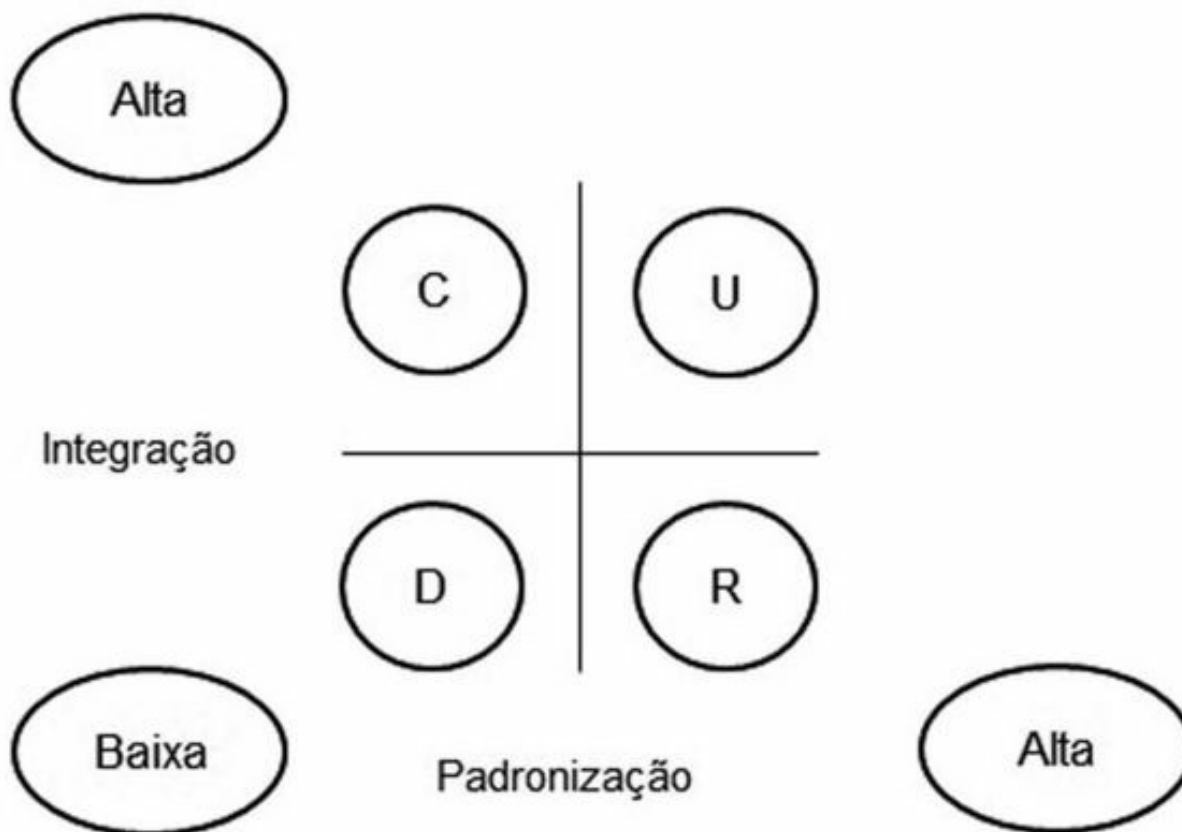


Figura 1-7 – Modelos Operacionais

Os quatro modelos sugeridos são resumidos a seguir:

- **Diversificação (D):** baixa padronização, baixa integração – envolve construir plataforma de serviços compartilhados que suportam entidades autônomas de negócio.
- **Coordenação (C):** baixa padronização, alta integração – envolve a construção de uma plataforma de informações compartilhadas para suportar decisões administrativas integradas.
- **Réplica (R):** alta padronização, baixa integração – envolve a construção de plataforma de tecnologias e processos de negócio padrão para definir uma marca comum.
- **Unificação (U):** alta padronização, alta integração – envolve a construção de uma plataforma de tecnologia, processos de negócios e dados compartilhados padronizados para suportar os requisitos globais de clientes de ponta a ponta.

A ideia central é que o negócio, em função da estratégia corrente, deve ter um modelo operacional que suporta esta estratégia. Por sua vez, o modelo operacional se traduz na forma de padronização e integração de processos. A arquitetura empresarial daria suporte ao modelo operacional. Os quatro exemplos citados demandariam arquiteturas empresariais diferentes. O modelo a ser perseguido seria o que refletisse melhor a estratégia do negócio.

Cabe ressaltar que o conceito de arquitetura empresarial (EA) não é o mesmo que o conceito de arquitetura de TI. Nem mesmo o conceito de arquitetura de processos pode ser considerado equivalente ao conceito de EA, que abrange a padronização e integração dos processos e se traduz pela “cola” entre estratégia e a execução da estratégia pela TI. EA é muito mais subjetivo do que a arquitetura de TI e trata da lógica de alto nível dos processos de negócios e capacidades da TI. EA é um conceito sofisticado e muitas vezes não entendido corretamente.

Assim, é preciso diferenciar claramente a Arquitetura Empresarial da Arquitetura de TI. Aqui se define arquitetura empresarial. No próximo tópico define-se o que seja arquitetura de TI.

A arquitetura empresarial é a lógica organizacional dos processos de negócio, dos dados, aplicativos e da infraestrutura de TI e deve refletir os requisitos de padronização e integração do modelo operacional da empresa. A arquitetura empresarial proporciona uma visão de longo prazo focada em construir capacidades e não só atender a demandas de curto prazo. Uma forma de entender o papel da arquitetura empresarial como fundação para vantagem competitiva é entender a visão baseada em recursos (*Resource Based View – VBR*).

O VBR é um modelo de desempenho com foco nos recursos e capacidades controlados por uma empresa como fontes de vantagem competitiva. Recursos no modelo VBR são definidos como ativos tangíveis e intangíveis que a empresa controla e que podem ser usados para criar e implementar estratégias. Capacidades são um subconjunto dos recursos de uma empresa e são definidas como ativos tangíveis e intangíveis que permitem a empresa aproveitar outros recursos que controla. Capacidades sozinhas não permitem que uma empresa crie e implemente suas estratégias. Capacidades são valiosas somente na medida em que permitem à empresa melhorar sua posição competitiva e visão individual dos funcionários e gerentes da empresa. Recursos organizacionais são atributos de grupos de pessoas e incluem sistemas formais e informais de planejamento.

O VBR se baseia na heterogeneidade e na imobilidade dos recursos. Em um ramo de atividade,

uma empresa pode ser mais competente do que outra, o que se convencionou chamar de heterogeneidade. Diferenças de recursos e capacidades de empresas podem ser duradouras, o que se convencionou chamar de imobilidade dos recursos. Estas suposições, consideradas juntas, permitem explicar por que algumas empresas superam outras, mesmo que estejam competindo no mesmo setor.

Recursos e Capacidades podem ser classificados em quatro categorias: financeiros, físicos, individuais e organizacionais. Recursos financeiros incluem o dinheiro que as empresas utilizam para criar e implementar estratégias. Recursos físicos incluem toda a tecnologia física utilizada em uma empresa. Isto inclui tecnologias de hardware e software. Recursos humanos incluem treinamento, experiência, inteligência e relacionamentos. Recursos Organizacionais são atributos de grupos e incluem estruturas, sistemas formais e informais de planejamento, cultura e até relações informais.

O que seriam capacidades de TI? São subconjuntos dos recursos de TI de uma empresa. Permitem que a empresa aproveite melhor processos de negócio – por exemplo, que implementam estratégias.

As capacidades de TI podem ser avaliadas segundo o modelo VRIO (*Value, Rarity, Imitability, Organization* – *Valor, Raridade, Imitabilidade, Organização*). O VRIO é utilizado para conduzir uma análise interna da organização. O VRIO sugere fazer quatro questões sobre capacidade ou recurso para determinar seu potencial competitivo:

- **A questão do valor:** o recurso permite que a empresa explore uma oportunidade e/ou neutralize uma ameaça?
- **A questão da raridade:** o recurso é atualmente controlado por apenas um pequeno número de empresas concorrentes?
- **A questão da imitabilidade:** as empresas sem esse recurso enfrentam problemas de custo para obtê-lo ou para desenvolvê-lo?
- **A questão da organização:** as outras políticas e os processos da empresa estão organizados para apoiar a exploração de seus recursos valiosos, raros e custosos de imitar?

A **Figura 1-8** ilustra a relação entre arquitetura empresarial e estratégia empresarial. Arquitetura Empresarial (também chamada de Plataforma Digitalizada) apoia o modelo operacional, que, por sua vez, permite executar a estratégia.

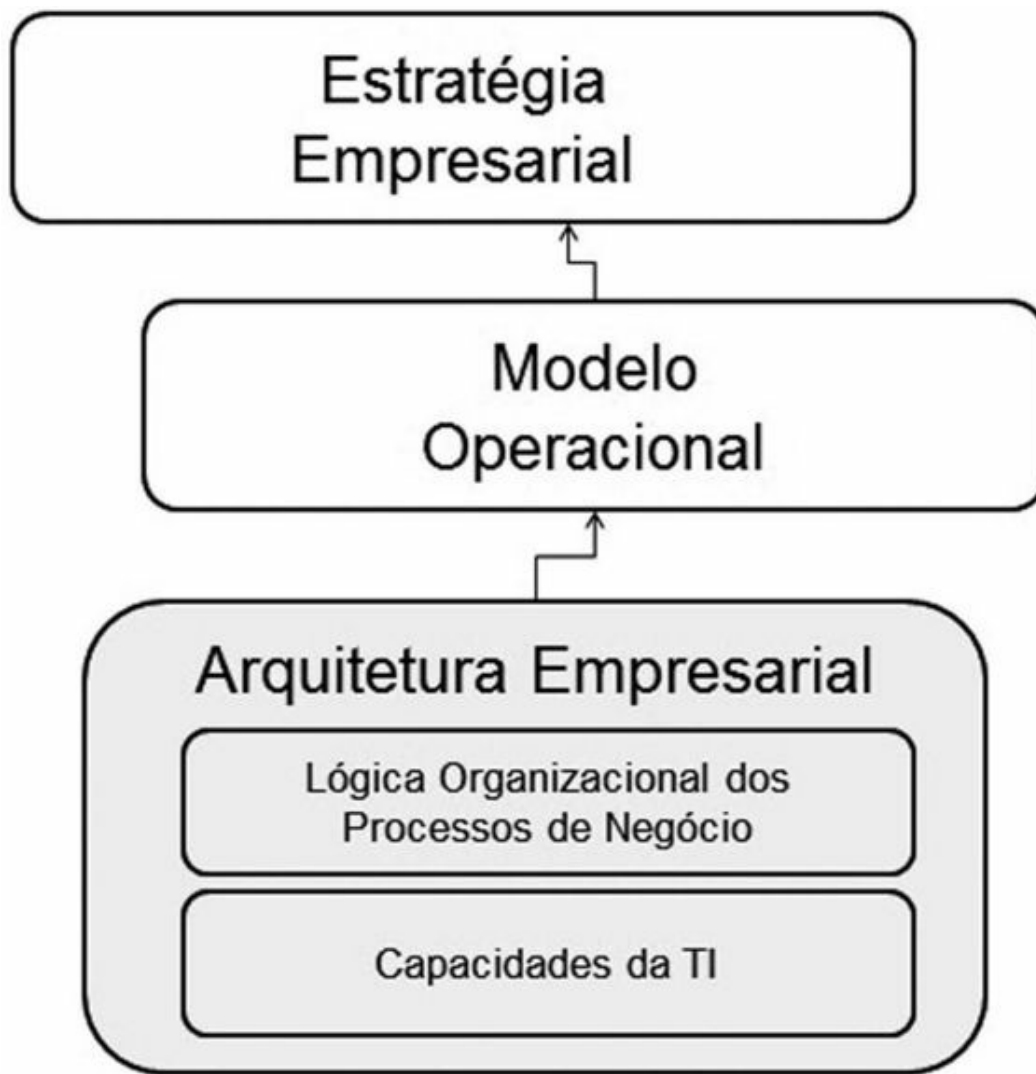


Figura 1-8 – Arquitetura Empresarial e Conceitos Relacionados

As empresas nos processos de construção da arquitetura empresarial normalmente avançam em estágios de aprendizado. São eles: Silos de Negócio, Tecnologia Padronizada, Núcleo Otimizado e Modularidade.

- **Arquitetura dos Silos de Negócio (Localização):** quando as empresas procuram maximizar as necessidades de cada unidade comercial ou as necessidades funcionais.
- **Arquitetura de Tecnologia Padronizada (Padronização):** a eficiência da TI é conseguida por meio da padronização tecnológica.
- **Arquitetura de Núcleo Otimizado (Otimização):** proporciona a padronização de dados e processos na empresa como um todo.
- **Arquitetura de Modularidade dos Negócios (Reutilização):** quando as empresas administram e reaproveitam componentes livremente associados aos processos de negócio habilitados pela TI com o intuito de preservar padrões globais e habilitar diferenças locais.

A ideia é que as empresas devem avançar no seu modelo de arquitetura empresarial passando por estágios. Como se aprendessem e assim fossem aprovadas para o estágio seguinte. Muitas

empresas consideram a mágica de tentar ter uma arquitetura modular de um momento para outro.

Em termos de arquitetura de TI, entende-se que esta arquitetura deve acompanhar o movimento da arquitetura empresarial, ou seja, arquitetura empresarial do tipo modularidade do negócio deve impor uma arquitetura de TI também modular.

Conforme constroem a EA, as empresas modificam o modelo de arquitetura de TI e o modelo de financiamento da TI.

Há uma boa chance de que a arquitetura modularizada, ou também chamada de reutilização, portanto, mais madura, tenha mais sucesso na concretização das metas estratégicas da organização. A arquitetura modular teoricamente possibilita maior agilidade estratégica. Em pesquisa realizada pelo MIT com grandes empresas americanas e europeias, poucas tinham atingido este estágio (cerca de 6%).

Os possíveis benefícios de adotar uma arquitetura empresarial seriam conseguir:

- Custos reduzidos de TI.
- Maior responsividade da TI.
- Melhor gestão do risco.
- Maior satisfação da administração.
- Melhores resultados de negócios.

A definição e a construção de uma arquitetura modular de negócios seriam mais importantes do que a simples execução do plano estratégico, considerando as grandes mudanças ambientais e a consequente necessidade de flexibilizar processos de negócio.

Os defensores do conceito de Arquitetura Empresarial defendem que a EA deve orientar também a terceirização da TI. O que terceirizar, o tipo de relacionamento com o provedor e os objetivos que realmente podem ser alcançados devem ser orientados pela maturidade da arquitetura empresarial (Ross, Weill, Robertson, 2006).

1.5. Arquitetura de TI

O conceito de Arquitetura Empresarial não é equivalente ao conceito de Arquitetura de TI, conforme já dito. São conceitos diferentes.

A arquitetura de TI deve ser pensada com base na Arquitetura Empresarial. Arquitetura Empresarial trata da lógica de alto nível dos processos de negócios e das capacidades da TI. Arquitetura de TI pode ser muitas coisas.

De forma geral, a arquitetura de TI trata das arquiteturas dos processos de negócio, das arquiteturas dos dados e informações, das arquiteturas das aplicações e da arquitetura da infraestrutura.

Considera-se como arquitetura de TI, para efeito de simplificação, as arquiteturas da

aplicação e da infraestrutura. A arquitetura de um DATACENTER seria uma espécie de subconjunto da arquitetura da infraestrutura.

A **Figura 1-9** relaciona as arquiteturas. É evidente que a arquitetura de TI deve suportar a arquitetura empresarial. Uma arquitetura empresarial do tipo reutilização não pode ter uma arquitetura de TI baseada em aplicativos e infraestrutura legada. Também é importante ressaltar que uma grande organização pode possuir arquiteturas empresariais em fases diferentes. Uma empresa pode até mesmo possuir modelos operacionais diferentes se existem unidades de negócio com diferentes níveis de maturidade.

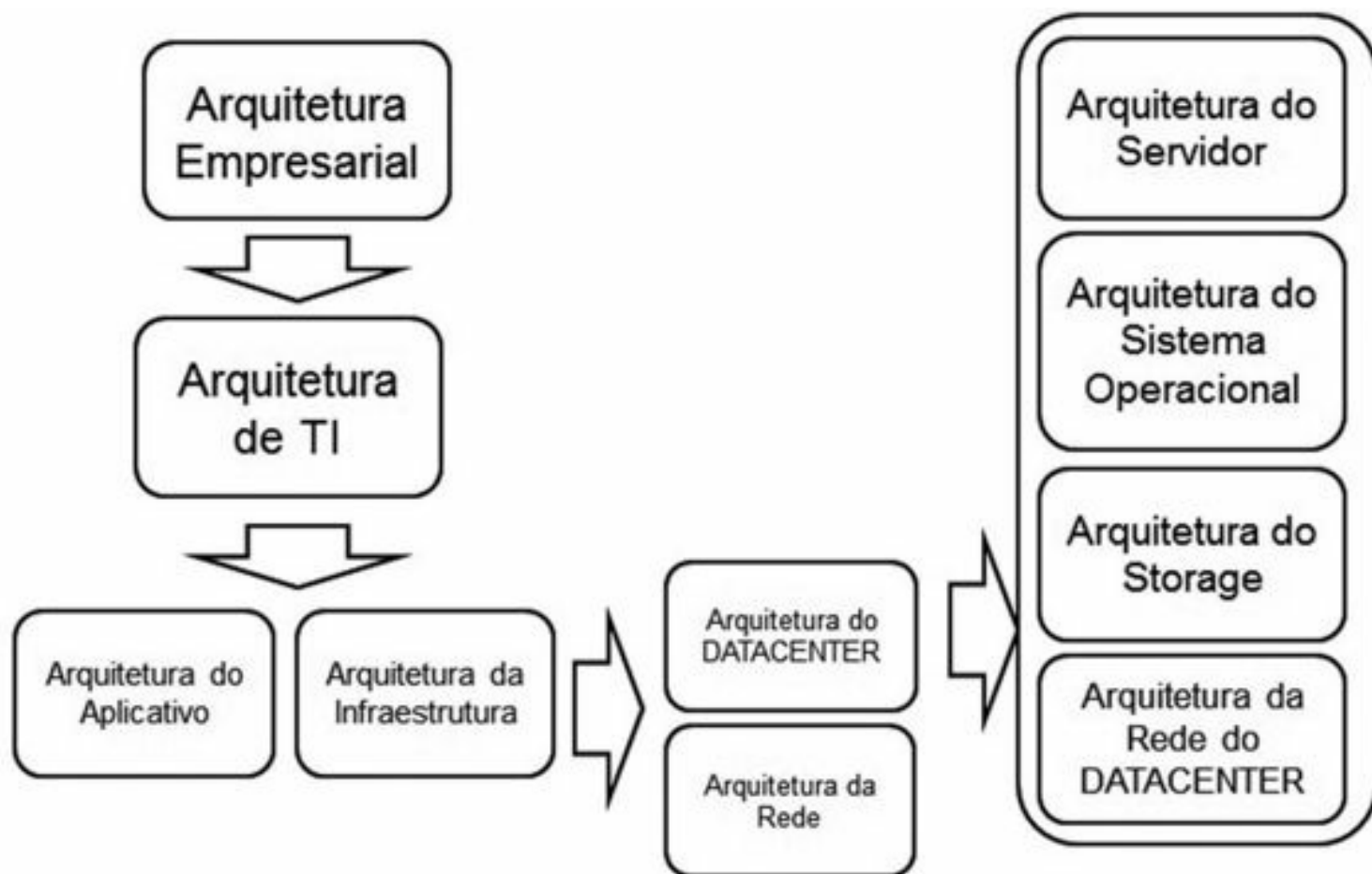


Figura 1-9 – Arquitetura Empresarial e Arquitetura de TI

Como evoluiu a arquitetura de TI? A arquitetura de TI inicialmente era centralizada (arquitetura MAINFRAME); depois se tornou descentralizada com a adoção do modelo cliente/servidor e agora volta a ser centralizada com a adoção da CLOUD COMPUTING. A **Figura 1-10** ilustra o avanço da arquitetura. Observa-se que as arquiteturas de TI, desde o início do seu uso comercial, aparecem como um movimento pendular: em certos momentos acontecem de forma centralizada; em outros, mais adiante, de forma descentralizada.

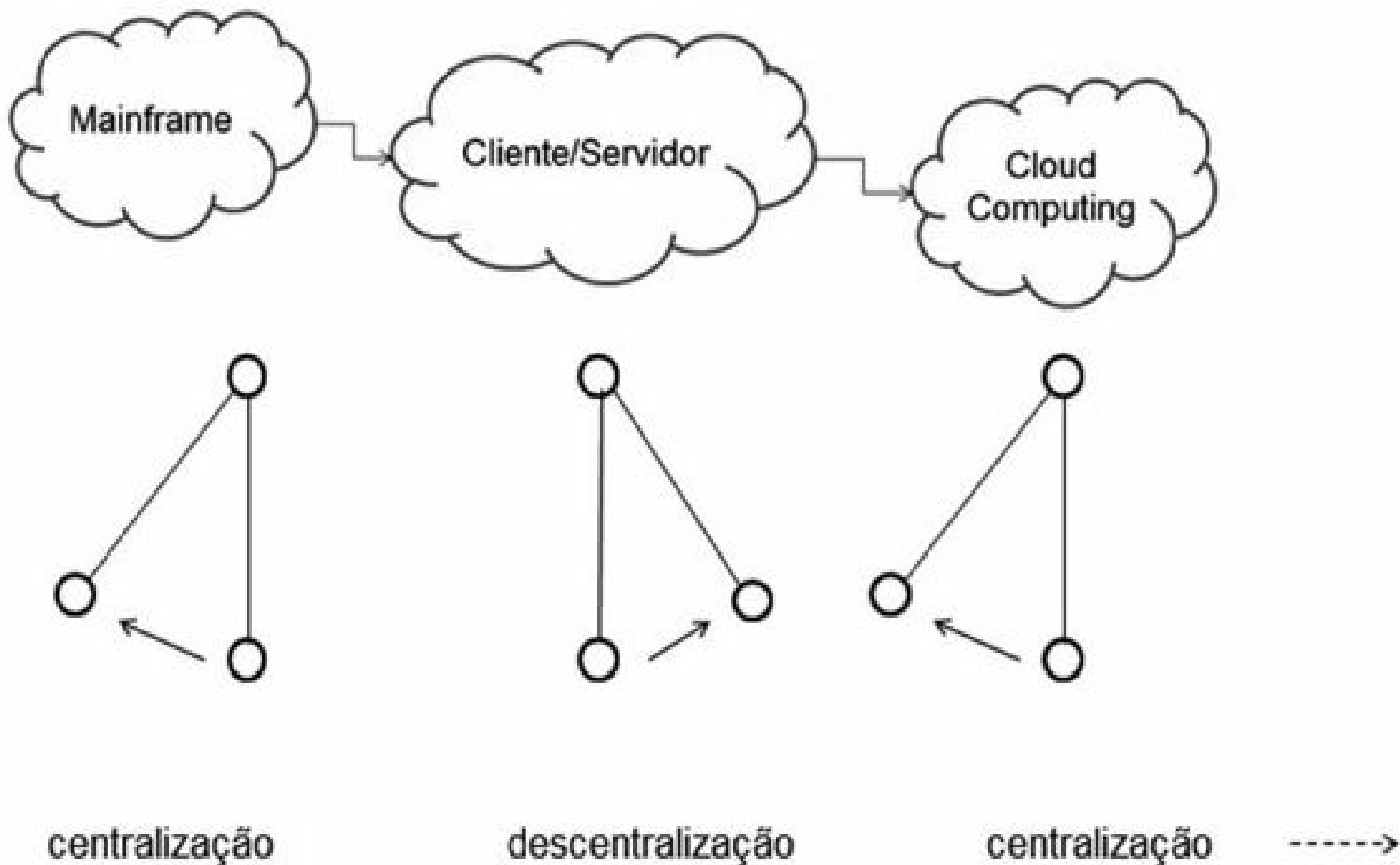


Figura 1-10 – Centralização e Descentralização da Arquitetura da TI

A arquitetura MAINFRAME, centralizada, era focada na melhor utilização dos recursos. Na época, a visão era de que o que deveria ser otimizado era a utilização dos recursos, que eram muito caros. Flexibilidade para o usuário ficava em segundo plano.

Com o surgimento do sistema operacional Windows e sistemas baseados em baixa plataforma e em rede, emergiu a arquitetura cliente/servidor, arquitetura distribuída, que trouxe flexibilidade, mas degradou a utilização dos recursos. Neste caso, os recursos eram mais baratos e perderam a importância quando comparados à flexibilidade introduzida. A boa utilização dos recursos ficava em segundo plano, considerando que com esta arquitetura o recurso era considerado barato.

A arquitetura CLOUD COMPUTING, a mais recente, baseada em grandes processadores e repositórios de dados, os DATACENTERS, é um meio termo entre as duas arquiteturas anteriores. Teoricamente, pode-se ter a otimização do uso dos recursos sem perda de flexibilidade. A arquitetura CLOUD COMPUTING é marcada por uma TI com grande crescimento da base de dados, mobilidade acentuada por parte dos usuários e diversas formas de dispositivos de acesso e abundante disponibilidade de aplicativos.

Também a arquitetura CLOUD COMPUTING, por suas características intrínsecas, permitiria tornar a TI mais flexível, o que seria uma condição importante para o repensar da TI. Os recursos seriam fornecidos sob demanda. Arquitetura Empresarial flexível demanda Arquitetura de TI modular.

A **Tabela 1-1** sintetiza as principais características das arquiteturas de TI.

Tabela 1-1 – Comparação entre as Arquiteturas de TI

	Tecnologia	Economia	Modelo de Negócio
MAINFRAME	Computação centralizada	Otimizado para eficiência por causa do alto custo	Alto custo de hardware e software
CLIENTE/SERVIDOR	Computação distribuída	Otimizado para agilidade devido ao baixo custo	Licença perpétua para SO e aplicativos
CLOUD COMPUTING	Grandes DATACENTERS	Otimizado para eficiência e agilidade	Paga pelo uso

Fonte: The Economics of the Cloud, Microsoft, Nov. 2010

1.6. Conceito na Prática: Oracle Enterprise Architecture Framework

A Oracle propõe um *framework* para EA, o *Oracle Enterprise Architecture Framework* (OEAF), mostrado na **Figura 1-11**. Para a Oracle, EA é um método e um princípio organizacional que alinha objetivos estratégicos com estratégia de TI e plano de execução. A EA, segundo a Oracle, propicia um guia para transformar a organização com a TI.

O *framework* OEAF da Oracle inclui:

- Vocabulário comum, modelos e taxonomia.
- Processos, princípios, estratégia e ferramentas.
- Arquitetura de referência e modelos.
- Guia prescritivo.
- Catálogo dos entregáveis de arquitetura e artefatos.
- *Enterprise Architecture Content Metamodel*.
- Recomendação para produtos e configurações (opcional).

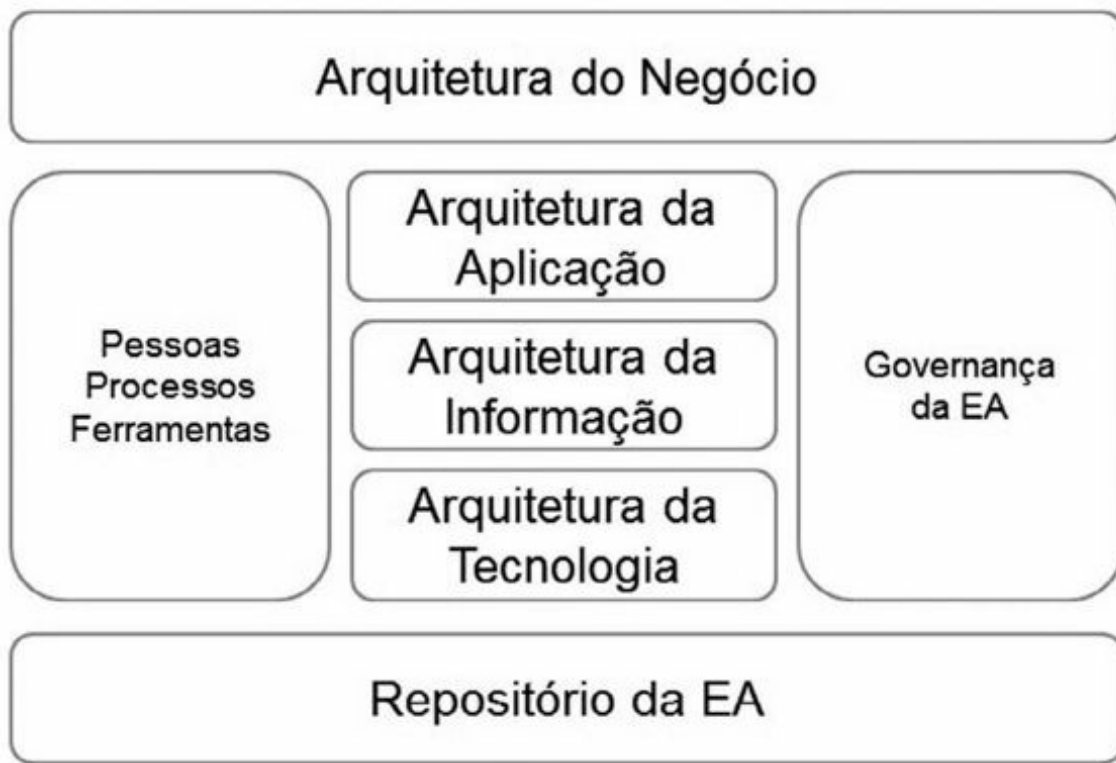


Figura 1-11 – Oracle Enterprise Architecture Framework (OEAF)

OEAF adiciona artefatos específicos da Oracle para EA, arquiteturas de referência, ferramentas e guias de melhores práticas de implementação e governança⁵.

Arquitetura do Negócio inclui:

- Estratégia do Negócio.
- Funções do Negócio.
- Organização do Negócio.

Arquitetura da Aplicação inclui:

- Estratégia da Aplicação.
- Serviços da Aplicação.
- Processos da Aplicação.
- Componentes Lógicos.
- Componentes Físicos.

Arquitetura da Informação inclui:

- Estratégia da Informação.
- Ativos de Informação.

Arquitetura da Tecnologia inclui:

- Estratégia da Arquitetura.
- Serviços de Tecnologia.
- Componentes Lógicos.
- Componentes Físicos.

Governança de EA inclui:

- Pessoas.
- Processos e Políticas.
- Tecnologia.
- Finanças.

Repositório de EA inclui:

- Artefatos de arquitetura e entregas que são capturados e desenvolvidos através do ciclo de vida da EA. Este repositório fornece informações sobre o estado atual da EA e uma biblioteca de arquiteturas de referência, modelos e princípios que descrevem o objetivo a ser atingido para a arquitetura.

Existem outros modelos de EA como o TOGAF, que está na versão 9, e o modelo do Gartner.

1.7. TI como Serviço (TlaaS) e CLOUD COMPUTING

O que seria um serviço? Serviço é um benefício que uma organização entrega para outra organização. No caso da TI, quando ela suporta o processo de negócio, ela está provendo um serviço para a organização.

Um serviço pode ser descrito por suas características, tais como:

- Agrega algum tipo de valor, e este valor gera seu controle e mensuração.
- É intangível – embora se perceba seus resultados, não se consegue materializar a execução do serviço e nem sempre se consegue mensurar o serviço, seus resultados e mesmo o benefício que ele traz.
- É produzido e consumido ao mesmo tempo.
- É produzido pela atuação organizada de um conjunto de processos, atuações, experiências e recursos de infraestrutura.

Um serviço envolve pelo menos três papéis:

- Um provedor responsável pela produção e entrega do serviço, sob políticas e ambiente diferentes do cliente.
- Um usuário que, de alguma forma, se beneficia diretamente com a entrega do serviço.

- Um cliente, que define a relação de funcionalidade e nível de serviço e o investimento no serviço. O cliente também percebe o benefício de um serviço, mas com um ponto de vista diferente do usuário.

Alguns outros aspectos importantes referentes a serviços que precisam ser mencionados: na prestação de serviços, deve sempre ser estabelecido um canal entre as partes provedoras e receptoras, pelo qual o serviço é executado e entregue. Normalmente, o canal afeta a qualidade da entrega do serviço. Como partes receptoras, os clientes e usuários veem o serviço como uma peça única e não separável em componentes. Também deve haver sempre algum retorno pela entrega de serviços. Este retorno pode ser transacional (pagamento por entrega), de orçamento (apropriação de item de custo) ou outros meios.

Existe certa confusão com os conceitos de TI orientada a serviços e a gestão de serviços propiciada pela TI. A TI pode gerenciar de forma eficiente os serviços oferecidos, mas pode não ter de fato uma orientação a serviços. Adotar a gestão dos serviços de TI baseada em melhores práticas como o *Information Technology Infrastructure Library* (ITIL) por si só não torna a TI um serviço, mas ajuda. ITIL é um conjunto de boas práticas a serem aplicadas na infraestrutura, operação e manutenção de serviços de TI. Foi desenvolvido no final dos anos 1980 pela CCTA (Central Computer and Telecommunications Agency) e atualmente está sob custódia da OGC (Office for Government Commerce) da Inglaterra. A ITIL está na versão 3.

O que seria então a TI como serviço TlaaS – *TI as a Service*? A TI passa a ser um negócio de serviços dentro do negócio. Basicamente, TlaaS trata de tornar a TI uma organização de serviços que cumpre acordos de serviços com clientes/usuários. A organização de TI estaria entregando um benefício, portanto um serviço, para outra organização que não é a de TI. Também este serviço seria consumido e produzido ao mesmo tempo. Se a organização demanda um melhor acordo de serviço, a ideia é que a TI rapidamente se altere e entregue o nível de serviço desejado.

TlaaS envolve a gestão da entrega de serviços, os processos de TI que possibilitam a entrega dos serviços, a arquitetura do conjunto de aplicativos orientada a serviços SOA – *Service Oriented Architecture*) e a infraestrutura também orientada a serviço SOI (Service Oriented Infrastructure). Ou seja, além de ter uma gestão orientada a serviços, as arquiteturas devem possibilitar a entrega destes serviços.

SOA estabelece uma plataforma de computação orientada a serviços. SOA se caracteriza por introduzir novas tecnologias e plataformas que suportam especificamente a criação, a execução e a evolução das soluções orientadas a serviços. SOA permite principalmente maior alinhamento do domínio de negócio e da tecnologia. Também maior interoperabilidade, mais opção de diversificação de fornecedores e maior retorno sobre o investimento.

No nível do aplicativo, os serviços fornecidos pela arquitetura SOA existem como softwares fisicamente independentes que dão suporte à obtenção dos objetivos estratégicos associados à computação orientada a serviços.

Computação orientada a serviço representa uma nova geração de plataforma de computação que abrange o paradigma da orientação a serviço e a arquitetura orientada a serviços.

SOI é a fundação para tornar a infraestrutura orientada a serviço. SOI tem sido adotado mais lentamente do que SOA. SOI facilita o reuso e a alocação dinâmica dos recursos necessários de infraestrutura. As características do serviço fornecem a base para o desenvolvimento dos serviços

como também para a disponibilização dos serviços de infraestrutura de TI. A ideia é que os serviços disponibilizados tenham a qualidade necessária e tenham comportamento consistente ao longo de todo o ciclo do serviço.

O modelo baseado em CLOUD COMPUTING, pelas características a serem vistas no Capítulo 2, permite que a TI de fato seja entregue como serviço ou pelo menos torne mais adequada a ideia da TI ser vista como serviço.

CLOUD COMPUTING permite desacoplar os processos de negócio da TI necessária para rodá-los. Ao mesmo tempo, introduz a ideia de elasticidade na utilização de infraestrutura. Recursos podem ser utilizados em períodos de alta demanda e devolvidos em períodos de baixa demanda. Também permite flexibilizar a alocação de custos para empresas que podem mudar de um modelo baseado em custo de capital para um modelo baseado em despesas operacionais.

CLOUD COMPUTING teoricamente possui escalabilidade infinita. Mas esta escalabilidade esbarra na arquitetura da aplicação e na infraestrutura disponível. Ou seja, a aplicação e a infraestrutura precisam ser escaláveis para a obtenção da escalabilidade desejada. Também a elasticidade é uma propriedade fundamental da CLOUD COMPUTING, e para que ela seja plenamente funcional a aplicação e a infraestrutura precisam ser construídas com uma arquitetura adequada.

Em resumo, o conceito de TI como um Serviço – inclui a entrega do software, a infraestrutura e as plataformas – oferece às organizações mais flexibilidade no uso do poder de TI para atender às necessidades comerciais e está associado à proposta da CLOUD COMPUTING.

A **Figura 1-12** ilustra a mudança da computação cliente/servidor para a CLOUD COMPUTING. CLOUD COMPUTING é um aprimoramento da orientação a serviço pregada pela arquitetura orientada a serviço (SOA) e pela infraestrutura orientada a serviço (SOI), que, por sua vez, já foi um avanço da arquitetura cliente/servidor. Os modelos PaaS e IaaS, típicos da arquitetura CLOUD COMPUTING, serão explicados posteriormente.

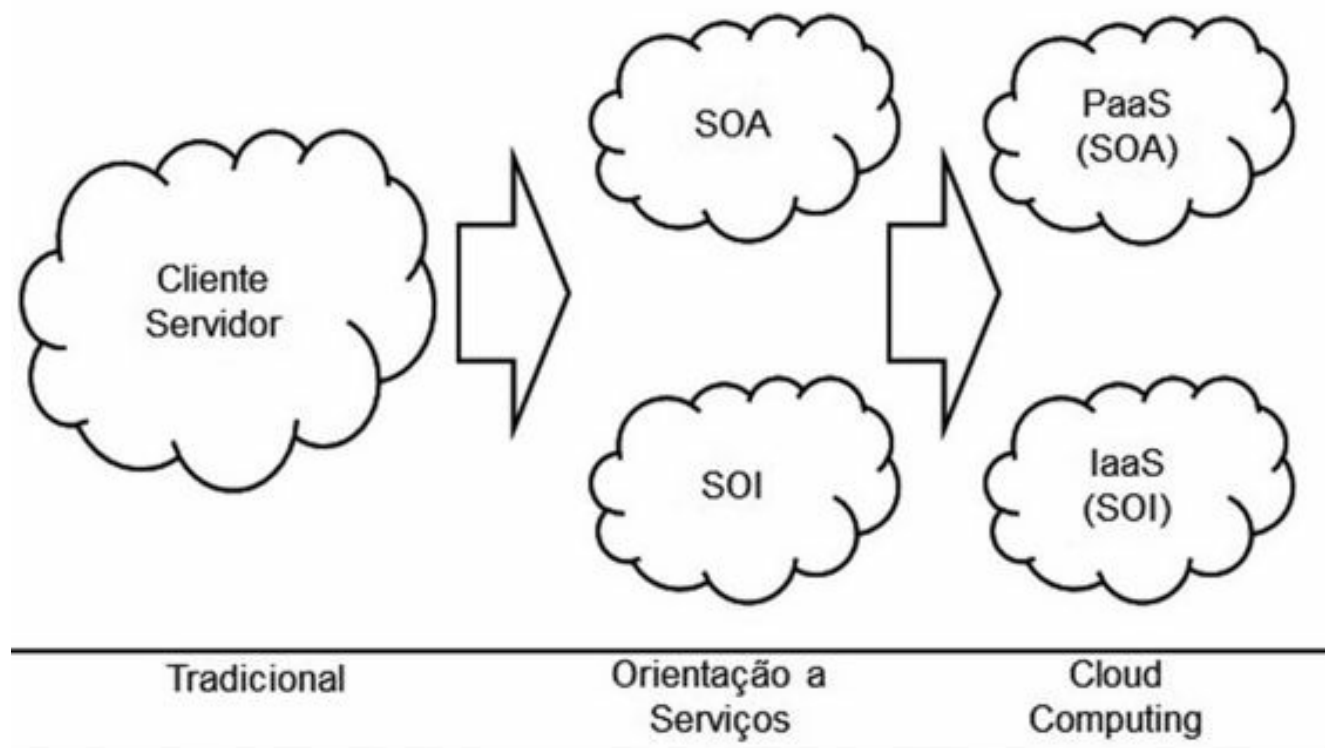


Figura 1-12 – Avanço da Arquitetura de TI

A busca de um novo modelo para TI passa por uma mudança na forma de financiar a TI e na arquitetura da TI que suporta o conceito de arquitetura empresarial. CLOUD COMPUTING é parte desta nova forma de pensar a TI.

1.8. Referências Bibliográficas

Architectural Strategies for CLOUD COMPUTING, Oracle, 2009.

Barney, Jay B. Strategic Management and Competitive Advantage: concepts, 2006, Pearson Education.

Big data: The next frontier for innovation, competition and productivity, McKinsey Global Institute, may 2011.

Carr, Nicholas. The big switch: rewiring the world, from Edison to Google. W. W. Norton Co, 2008.

Erl, Thomas. SOA principles of service design, SOA Systems, Inc. 2008.

IT as a Service: Transforming IT with the Windows Azure Platform, Microsoft, versão 1.0, november 2010.

Reimaging IT: The 2011 CIO Agenda, Gartner.

Ross, Jeanne W.; Weill, Peter; Robertson, David C. Enterprise Architecture as Strategy. Harvard Business School Publishing, 2006.

Ross, Jeanne, Weill, Peter. IT Governance – How Top performers manage IT decisions Right for Superior Results, Harvard Business School Publishing, 2004.

The Economics of the Cloud, Microsoft, november. 2010.

Weill, Peter; Ross, Jeanne. IT Savvy, What Top Executives Must Know To Go From Pain To Gain.
Harvard Business School Publishing, 2009.

2. Visão Geral

2.1. Introdução

As empresas estão se organizando em formato de rede. Os processos de negócio entre estas organizações se utilizam cada vez mais dos aplicativos que processam e fornecem as informações necessárias para o pleno funcionamento deste novo arranjo. Por sua vez, a TI é a “cola” que permite que estas organizações trabalhem em conjunto com um determinado objetivo e entregue para o cliente final um valor que, somado, é maior do que o que se conseguiria com as partes isoladas sem a TI. Ou seja, a TI é que viabiliza a organização em rede.

Esta nova organização, em rede, diferente da empresa vertical de Ford, é baseada em arranjos horizontais. A ideia da rede é ser uma opção para mercados e hierarquia e o formato em rede se beneficia da possível redução dos custos de transação propiciado pela TI.

A **Figura 2-1** ilustra a referida mudança. A TI, no caso da organização com forma de pirâmide, era de certa forma centralizada e exclusiva, mesmo que os serviços fossem prestados por terceiros.

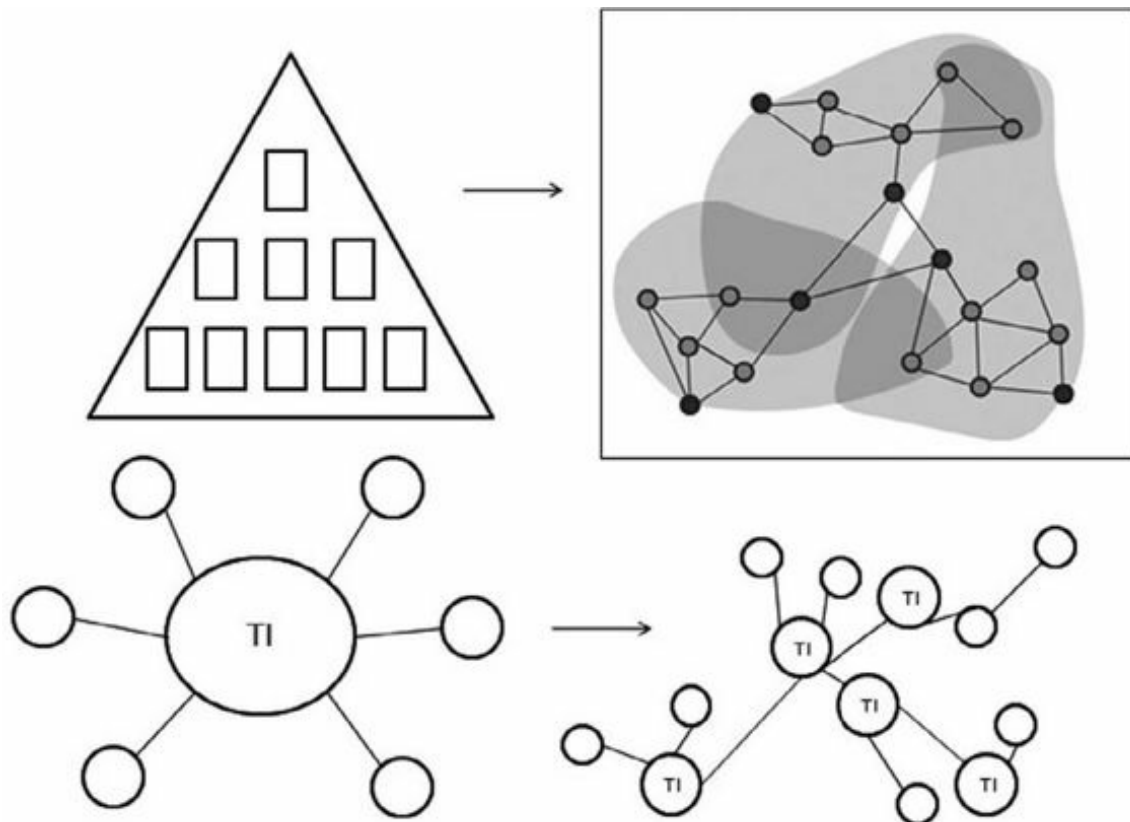


Figura 2-1 – Mudança Organizacional alicerçada pela TI

A arquitetura utilizada era baseada em grandes computadores, MAINFRAMES, que eram muito caros. A preocupação era a de otimizar o uso deste recurso. A flexibilidade para os usuários ficava em segundo plano.

Os computadores pessoais (PCs) surgiram na década de 80 e não foram vistos inicialmente como ferramentas para utilização em negócios. Com o surgimento do Windows e das redes locais e a possibilidade de ligar os PCs a servidores de rede, iniciou-se uma nova forma de computação empresarial, a computação cliente/servidor. A computação cliente/servidor era ineficaz, mas, com o aumento da capacidade dos processadores, o barateamento destes mesmos processadores e uma oferta abundante de aplicações, tornou-se hegemônica.

A arquitetura padrão passou a ser cliente/servidor. Os aplicativos rodavam parte no servidor e parte na estação cliente, distribuindo o processamento. Servidores e clientes eram conectados por uma rede que utilizava protocolos não padronizados. Os recursos utilizados nesta nova configuração eram baratos, conforme dito anteriormente, e o que se priorizou foi a flexibilidade em detrimento da boa utilização dos recursos. Com o aumento do poder de processamento dos processadores e a redução do custo ocorrida com os grandes ganhos de escala, os recursos computacionais foram cada vez mais adquiridos e subutilizados. Os aplicativos obrigatoriamente rodavam em servidores distintos, o que tornava a utilização dos recursos cada vez mais ineficiente. A VIRTUALIZAÇÃO dos recursos veio resolver este aspecto, mas surgiu para uso em baixa plataforma, só muito tempo depois. A arquitetura distribuída provocou problemas no gerenciamento do ambiente de TI e trouxe muito descontrole, minando o baixo custo dos servidores.

A rede que dava suporte à computação cliente/servidor inicialmente era interna, ou seja, lá trafegavam dados e informações de uma mesma organização. Formar redes entre organizações era uma tarefa complexa e cara. Existiam quase sempre intermediários que comandavam a interligação entre as redes e cobravam caro por isto. Além disso, os padrões e protocolos utilizados eram muito específicos de cada setor.

O surgimento da Internet, a grande rede, e a consequente redução dos custos de interligação, o avanço da padronização de protocolos de comunicação com a adoção do conjunto de protocolos TCP/IP e a disponibilidade de banda que ocorreram muito rapidamente, um efeito positivo do exagero com os negócios na Internet, por volta do ano 2000, possibilitou tornar a opção de interligar redes de diferentes organizações uma realidade. Sistemas entre organizações eram agora baseados nestas redes que utilizavam a Internet como espinha dorsal.

Nesta fase, a Internet, para muitas organizações, era simplesmente um meio barato de ligação entre servidores e clientes. Com a popularização da Internet e a redução dos custos de conexão, a rede passou a ter uma abrangência maior. A utilização da Internet avançou rapidamente. Provedores de acesso surgiram em praticamente todos os locais do mundo de forma muito rápida.

Agora são tantos pontos de conexão que a figura da nuvem (CLOUD) para representar a TI parece ser mais adequada, conforme ilustra a **Figura 2-2**. Passou-se a considerar processar e armazenar os dados corporativos na própria rede. Esta é a ideia central de utilizar a TI como serviço público.

O que seria CLOUD COMPUTING? Na verdade, o conceito *ainda se aprimora*. A ideia inicial da CLOUD COMPUTING foi processar as aplicações e armazenar os dados fora do ambiente corporativo, dentro da grande rede, em estruturas conhecidas como DATACENTERS, otimizando o

uso dos recursos. DATACENTERS, em resumo, processam aplicações e armazenam os dados de organizações que atuam em rede.

O conceito de nuvem colocado antes é o conceito atual de nuvem pública (PUBLIC CLOUD). A ideia central da nuvem pública é permitir que as organizações executem boa parte dos serviços que hoje são executados em DATACENTERS corporativos em DATACENTERS na rede, providos por terceiros, podendo sair de um modelo baseado em Capex (custo de capital) para um modelo baseado em Opex (custo de operação) e onde agora os indicadores de desempenho estão atrelados aos níveis de serviço, principalmente disponibilidade e desempenho, acordados entre clientes e provedores. Estes acordos idealmente deveriam variar em função da criticidade da aplicação, o que na prática é um desafio.

Verifica-se que o conceito de CLOUD COMPUTING hoje é mais abrangente. Pode-se inclusive ter uma nuvem privada, cujos dados e aplicações fazem parte de uma única organização, ou mesmo uma nuvem híbrida, formada por nuvens públicas e privadas. As nuvens privadas ou mesmo híbridas deixam de ter algumas características encontradas em um modelo de nuvem pública, pelo menos inicialmente, mas possibilitam que o cliente tenha maior controle sobre a infraestrutura utilizada.

A arquitetura CLOUD COMPUTING significa mudar fundamentalmente a forma de operar a TI, saindo de um modelo baseado em aquisição de equipamentos para um modelo baseado em aquisição de serviços. A CLOUD COMPUTING, com a VIRTUALIZAÇÃO, teoricamente permite obter o melhor dos mundos: otimização do uso dos recursos e flexibilidade para o usuário.

Atualmente os serviços de CLOUD COMPUTING decorrentes da organização em nuvem pública são fornecidos em sua grande maioria por grandes organizações como Google, Microsoft, Amazon, Rackspace e outros grandes provedores regionais que hospedam e executam os aplicativos dos clientes empresariais. A tendência é que esta oferta cresça e mais opções apareçam, aumentando a concorrência nesta área e reduzindo os custos de utilização destes serviços.

A nova arquitetura introduzida pela CLOUD COMPUTING permite que as organizações escolham o modelo adequado para a arquitetura dos seus aplicativos e onde armazenar os seus dados. Isto inclui aplicativos que rodam internamente (*on-premise*), serviços públicos de nuvem e/ou serviços privados de nuvem.

O elemento central do processamento e armazenamento dos dados e da informação na nuvem é o DATACENTER (DC), conforme ilustra a **Figura 2-2**. Ou seja, agora a TI é novamente centralizada em grandes pontos de armazenamento e processamento (otimiza o uso dos recursos), os tais DATACENTERS, mas conserva a estrutura de interligação em redes (flexibilidade). A nuvem na verdade é um conjunto de grandes pontos de armazenamento e processamento de dados e informação. Outras características citadas a seguir reforçam o que se convencionou chamar de CLOUD COMPUTING.

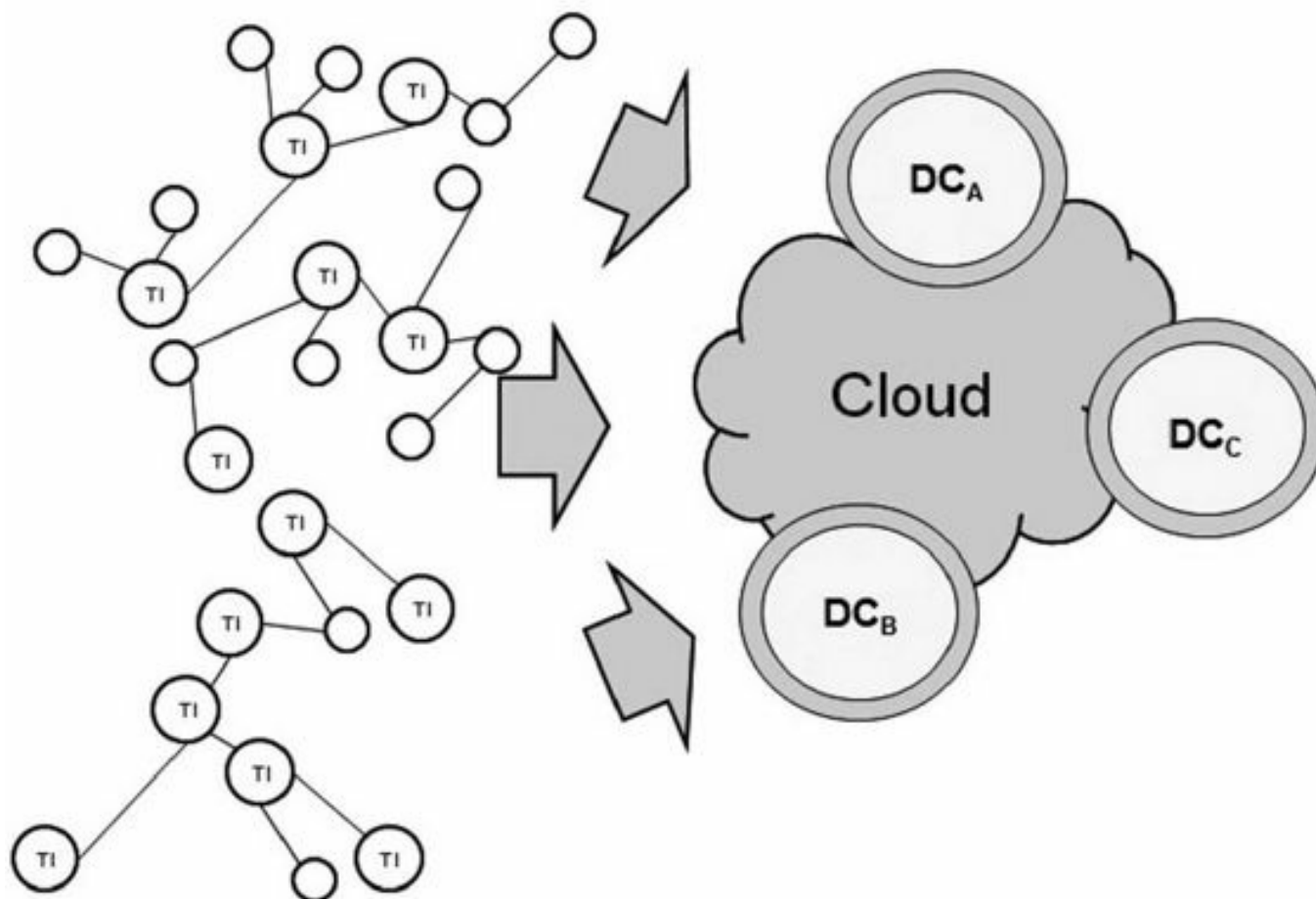


Figura 2-2 – Formação da CLOUD de TI

A centralização ou consolidação em grandes estruturas, como os DATACENTERS na arquitetura CLOUD COMPUTING, é viabilizada pela atual oferta de banda e pela existência de tecnologias que permitem alto poder de processamento e armazenamento em estruturas que reduzem o custo total de propriedade (TCO) da infraestrutura de TI. Estas estruturas, por outro lado, demandam muita energia e consequente refrigeração, tornando os projetos mais complexos.

A **Figura 2-3** ilustra a relação entre CLOUD COMPUTING, DATACENTERS e VIRTUALIZAÇÃO. CLOUD COMPUTING é formada por vários DATACENTERS e, por sua vez, os DATACENTERS são formados por diversos pools de recursos virtuais (PRV na **Figura 2-3**). Estes pools de recursos podem inclusive envolver mais de um DATACENTER.

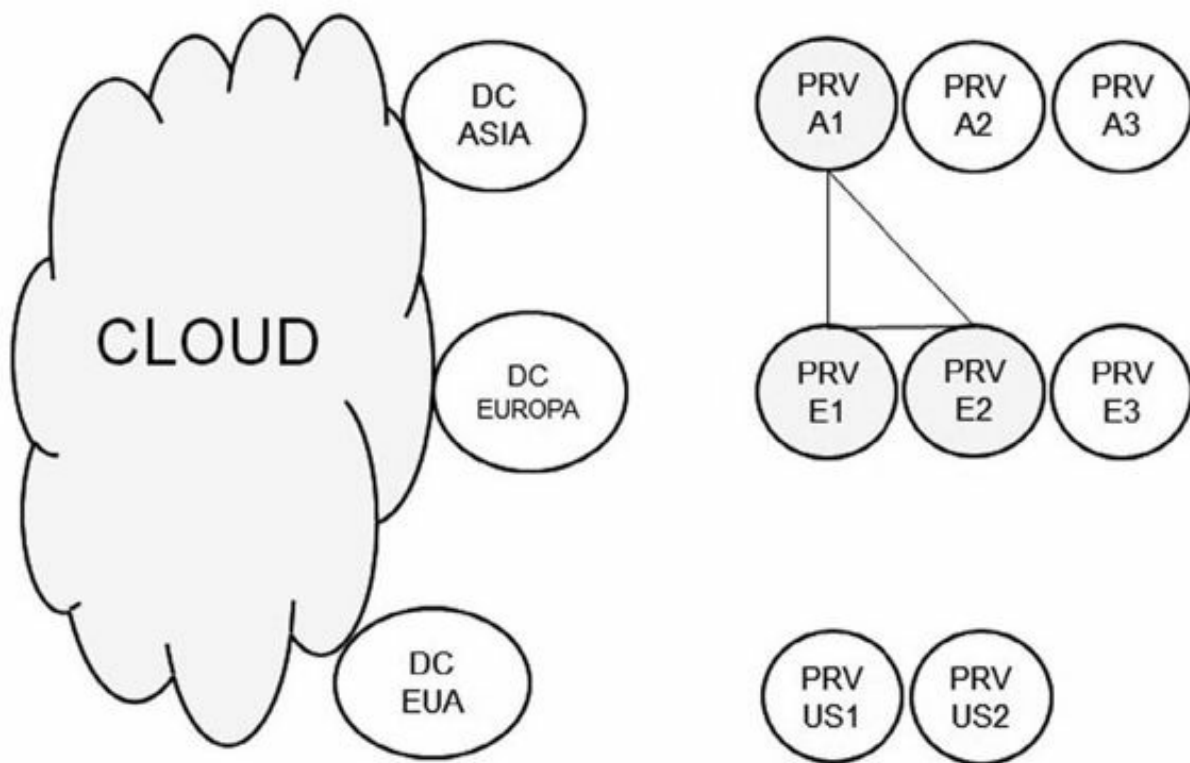


Figura 2-3 – Relação entre CLOUD COMPUTING, DATACENTER e VIRTUALIZAÇÃO

É como se as organizações agora aproveitassem o melhor dos mundos. Centralizam o processamento e o armazenamento da informação em grandes estruturas que propiciam ganho de escala, ao mesmo tempo em que estão integradas em rede, permitindo obter flexibilidade.

Em resumo, a centralização da base de informação em grandes DATACENTERS e a implantação de políticas de infraestrutura de rede, incluindo a segurança dos dados, que permitam o fornecimento de níveis adequados de serviço para as empresas clientes, providos por terceiros, é o panorama que norteia a TI.

O tamanho dos DATACENTERS é outra questão relevante. Qual seria o tamanho adequado? O ganho de escala e as novas tecnologias que permitem melhorar a eficiência energética têm possibilitado a construção de DATACENTERS de grandes dimensões. A [Figura 2-4](#) ilustra o novo DATACENTER do Facebook, em Oregon, nos Estados Unidos.



Figura 2-4 – DATACENTER do Facebook

O tamanho ideal não se sabe exatamente. O que se sabe é que as organizações buscam freneticamente aprender sobre a construção destas estruturas e otimizar a sua eficiência energética.

CLOUD COMPUTING faz sentido? Segundo a IBM, 85% da capacidade de computação do mundo está ociosa. CLOUD COMPUTING é parte da evolução natural da TI e um meio para aprimorar o uso da capacidade computacional em todo o mundo. Diferente do MAINFRAME, também é flexível no uso. As organizações podem inclusive contratar só parte de serviços de nuvem.

2.2. Conceito

CLOUD COMPUTING pode ser definido de várias formas. Apresentam-se aqui os conceitos mais bem aceitos e que possuem aplicabilidade.

CLOUD COMPUTING é um conjunto de recursos virtuais facilmente utilizáveis e acessíveis, tais como hardware, software, plataformas de desenvolvimento e serviços. Estes recursos podem ser dinamicamente reconfigurados para se ajustarem a uma carga de trabalho (WORKLOAD) variável, permitindo a otimização do seu uso. Este conjunto de recursos é tipicamente explorado através de um modelo *pague-pelo-uso*, com garantias oferecidas pelo provedor através de acordos de nível de serviços (Vaquero et al, 2009).

CLOUD COMPUTING é substituir ativos de TI que precisam ser gerenciados internamente por funcionalidades e serviços do tipo *pague-conforme-crescer* a preços de mercado. Estas funcionalidades e serviços são desenvolvidos utilizando novas tecnologias como a VIRTUALIZAÇÃO, arquiteturas de aplicação e infraestrutura orientadas a serviço e tecnologias e protocolos baseados na Internet como meio de reduzir os custos de hardware e software usados para

processamento, armazenamento e rede.

A **Figura 2-5** ilustra a mudança do modelo, onde a capacidade do sistema acompanharia a demanda, promovendo a otimização do uso dos recursos. A **Figura 2-5 (a)** ilustra a situação comum, onde a capacidade está sempre sobrando, e a **Figura 2-5 (b)** ilustra a situação ideal.

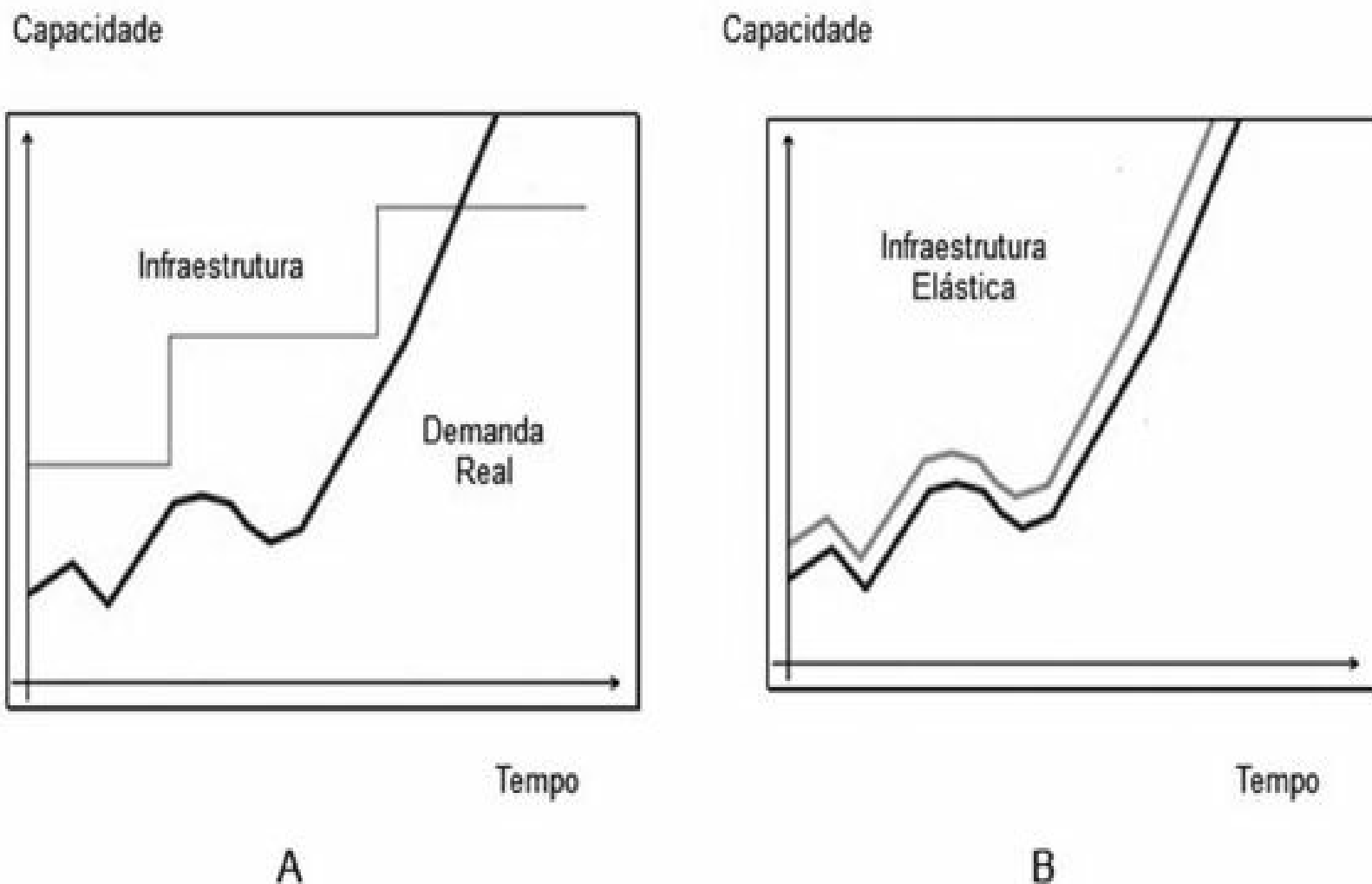


Figura 2-5 – Mudança com CLOUD COMPUTING

O benefício da elasticidade da CLOUD COMPUTING permite transferir o risco da baixa utilização e da alta utilização (saturação) para uma situação de ajuste fino entre a carga de trabalho e os recursos disponíveis. A ideia central da CLOUD COMPUTING é possibilitar que as aplicações que rodam em DATACENTERS isolados possam rodar na nuvem em um ambiente de larga escala e de uso elástico de recursos.

Elasticidade tem a ver com a demanda atual e a demanda futura. Com o uso maciço da Internet, as organizações já não sabem exatamente como os clientes se comportam. Imaginem uma companhia aérea vendendo bilhete a um preço irrisório no fim de semana e voltando à operação normal na segunda-feira. O que fazer com a infraestrutura de TI? No fim de semana precisa-se de muito recurso, mas já na segunda-feira os recursos estariam sobrando. Esta característica é difícil de ser obtida quando a organização utiliza uma infraestrutura interna. A VIRTUALIZAÇÃO utilizada internamente ajuda, mas pode não resolver. Como solicitar banda à operadora para o fim de semana e depois

devolver esta mesma banda de forma rápida? Existem questões ainda difíceis de serem resolvidas.

A proposta da CLOUD COMPUTING é de alguma forma melhorar o uso dos recursos e tornar a operação de TI mais econômica. Também é evidente que só a elasticidade propiciada pelos componentes da nuvem, os DATACENTERS, não é suficiente para garantir a elasticidade requerida pelas organizações no uso dos recursos. Também os aplicativos e as suas arquiteturas deverão ser orientadas a serviço para garantir a elasticidade. Tradicionais *approaches* baseados em técnicas de escalabilidade horizontal (*scale-out*) e escalabilidade vertical (*scale-up*) utilizadas tanto para o hardware como para o software em muitos casos já não são mais suficientes.

Em janeiro de 2011 o NIST (*National Institute of Standards and Technology*) publicou seu *draft* especial que definiu CLOUD COMPUTING em torno de suas cinco principais características, três modelos de serviço e quatro modelos de implantação. Esta forma de definir CLOUD COMPUTING é a mais bem aceita da atualidade e a mais abrangente e será tratada nos próximos itens.

2.3. Características Essenciais

Segundo o NIST, um modelo de CLOUD COMPUTING deve apresentar algumas características essenciais descritas a seguir:

- **Autoatendimento sob demanda:** funcionalidades computacionais são providas automaticamente sem a interação humana com o provedor de serviço.
- **Ampla acesso a serviços de rede:** recursos computacionais estão disponíveis através da Internet e são acessados via mecanismos padronizados, para que possam ser utilizados por dispositivos móveis e portáteis, computadores, etc.
- **Pool de recursos:** recursos computacionais (físicos ou virtuais) do provedor são utilizados para servir a múltiplos usuários, sendo alocados e realocados dinamicamente conforme a demanda.
- **Elasticidade rápida:** as funcionalidades computacionais devem ser rápidas e elasticamente providas, assim como rapidamente liberadas. O usuário dos recursos deve ter a impressão de que ele possui recursos ilimitados, que podem ser adquiridos (comprados) em qualquer quantidade e a qualquer momento. Elasticidade tem três principais componentes: escalabilidade linear, utilização on-demand e pagamento por unidades consumidas de em recurso.
- **Serviços mensuráveis:** os sistemas de gerenciamento utilizados pela CLOUD COMPUTING controlam e monitoram automaticamente os recursos para cada tipo de serviço (armazenamento, processamento e largura de banda). Esse monitoramento do uso dos recursos deve ser transparente para o provedor de serviços, assim como para o consumidor do serviço utilizado.

A **Figura 2-6** resume as características essenciais da nuvem definidas pelo NIST.

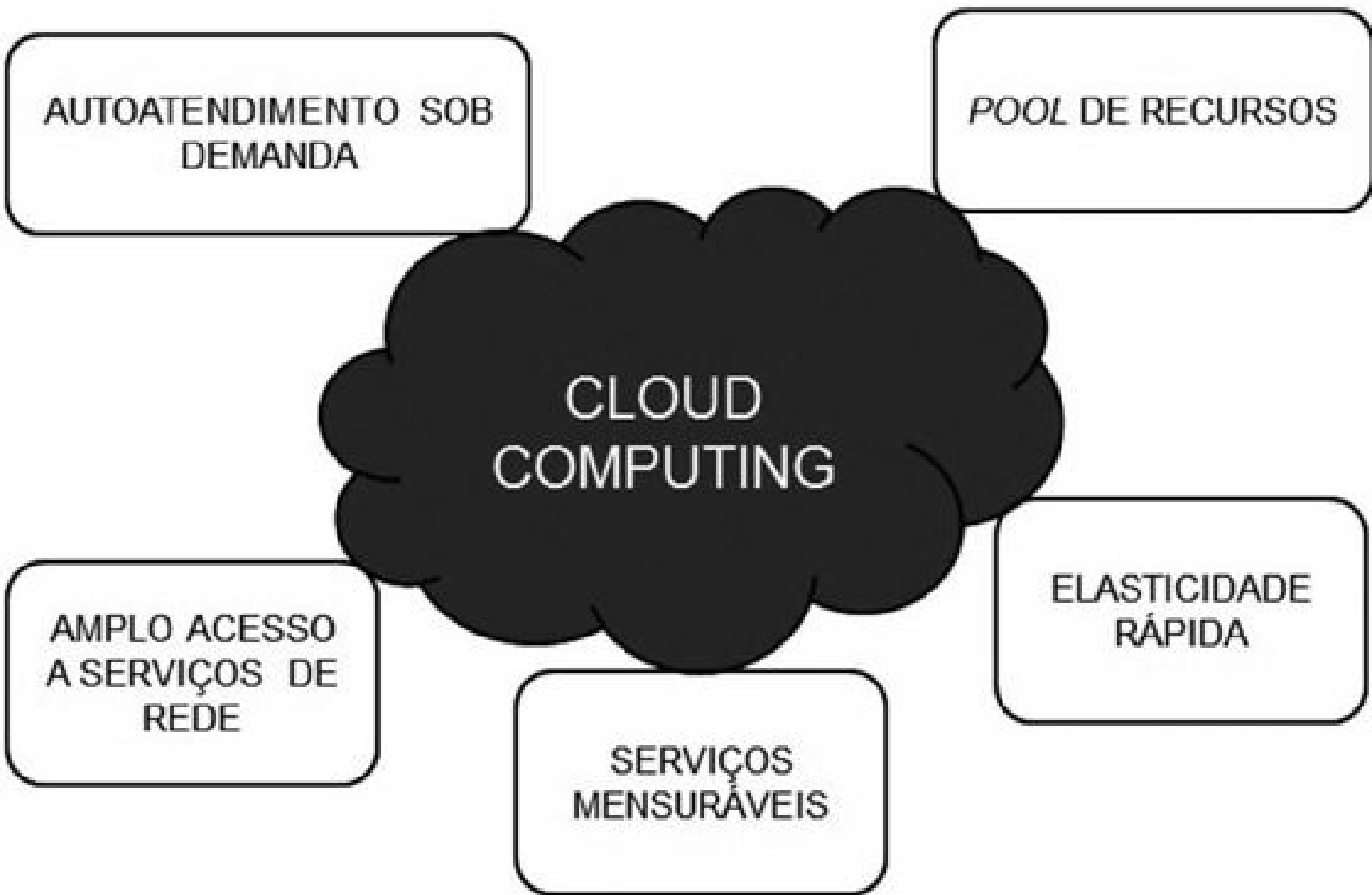


Figura 2-6 – Características da CLOUD COMPUTING

A proposta da CLOUD COMPUTING é criar a ilusão de que o recurso computacional é infinito e ao mesmo tempo permitir a eliminação do comprometimento antecipado da capacidade. Além disso, a ideia é permitir o pagamento pelo uso real dos recursos. Do ponto de vista prático, os provedores, em sua grande maioria, ainda não estão preparados para disponibilizar esta forma de serviço e muitos ainda têm dificuldade de gerenciar os níveis de serviço acordados.

O modelo da Amazon, o Amazon AWS, é na atualidade o que melhor descreve o modelo conceitual de CLOUD COMPUTING do NIST.

2.4. Modelos de Serviço

Existem três principais modelos de serviços para CLOUD COMPUTING:

- **Infraestrutura como um serviço (*Infrastructure as a Service – IaaS*):** capacidade que o provedor tem de oferecer uma infraestrutura de processamento e armazenamento de forma transparente. Neste cenário, o usuário não tem o controle da infraestrutura física, mas, através de mecanismos de VIRTUALIZAÇÃO, possui controle sobre as máquinas virtuais, armazenamento, aplicativos instalados e possivelmente um controle limitado dos recursos de rede. Um exemplo de IaaS é a opção Amazon EC2.

- **Plataforma como um serviço (*Platform as a Service – PaaS*):** capacidade oferecida pelo provedor para o desenvolvedor de aplicativos que serão executados e disponibilizados na nuvem. A plataforma na nuvem oferece um modelo de computação, armazenamento e comunicação para os aplicativos. Exemplos de PaaS são a AppEngine do Google e o Windows Azure da Microsoft.
- **Software como um serviço (*Software as a Service – SaaS*):** aplicativos de interesse para uma grande quantidade de clientes passam a ser hospedados na nuvem como uma alternativa ao processamento local. Os aplicativos são oferecidos como serviços por provedores e acessados pelos clientes por aplicações como o browser. Todo o controle e gerenciamento da rede, sistemas operacionais, servidores e armazenamento é feito pelo provedor de serviço. O Google Apps e o [SalesForce.com](https://www.salesforce.com) são exemplos de SaaS.

A **Figura 2-7** ilustra as possíveis arquiteturas para CLOUD COMPUTING quando comparadas aos serviços tradicionais de DATACENTER conhecidos como *Hosting* e *Colocation*.

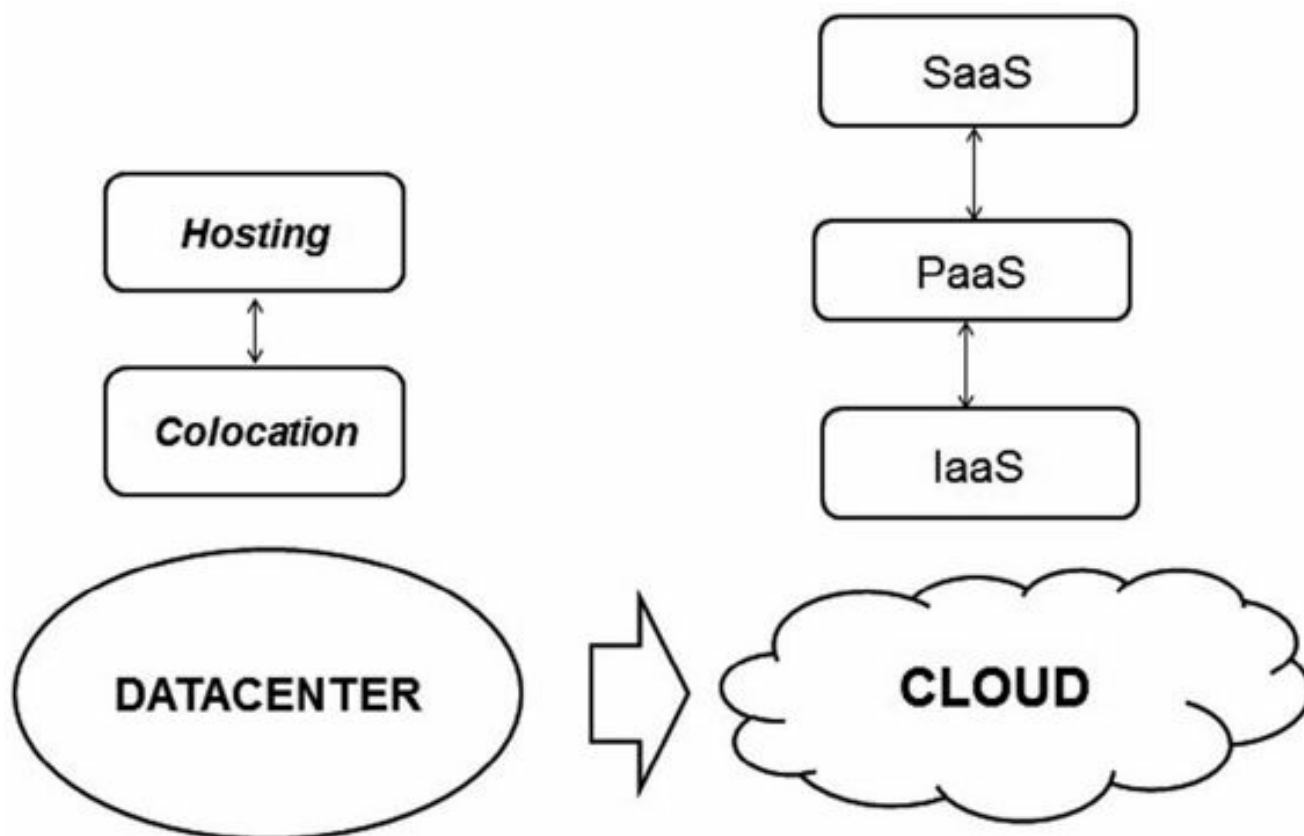


Figura 2-7 – Mudança no perfil dos serviços

- **Colocation:** a organização contrata o espaço físico dos RACKS e a infraestrutura de energia e de telecomunicações, porém os servidores, as aplicações, o gerenciamento, o monitoramento e o suporte técnico são fornecidos pela organização contratante. Esta relação pode ser flexibilizada e para isto costuma-se estabelecer um contrato com os termos e as condições, definindo claramente o escopo dos serviços de cada lado.
- **Hosting:** oferece uma linha de serviços indicada para organizações que desejam

aperfeiçoar investimentos em hardware e software. O serviço de *hosting* permite à organização contratante a utilização da infraestrutura do DATACENTER, incluindo servidores, *storage* e unidade de backup, além de contar com os profissionais do provedor de serviços para suporte.

Assim, deve ficar claro que CLOUD COMPUTING é uma proposta muito mais sofisticada do que a simples oferta de serviços de *Hosting* e *Colocation*. Alguns provedores de Internet confundem os clientes com o anúncio de disponibilidade de serviços de nuvem, mas essencialmente operam no modelo anterior.

Para entender melhor CLOUD COMPUTING, pode-se tentar identificar os papéis desempenhados na arquitetura baseada em nuvem. A **Figura 2-8** destaca quem fornece serviços e quem consome. O modelo IaaS suporta o modelo PaaS, que suporta o modelo SaaS.

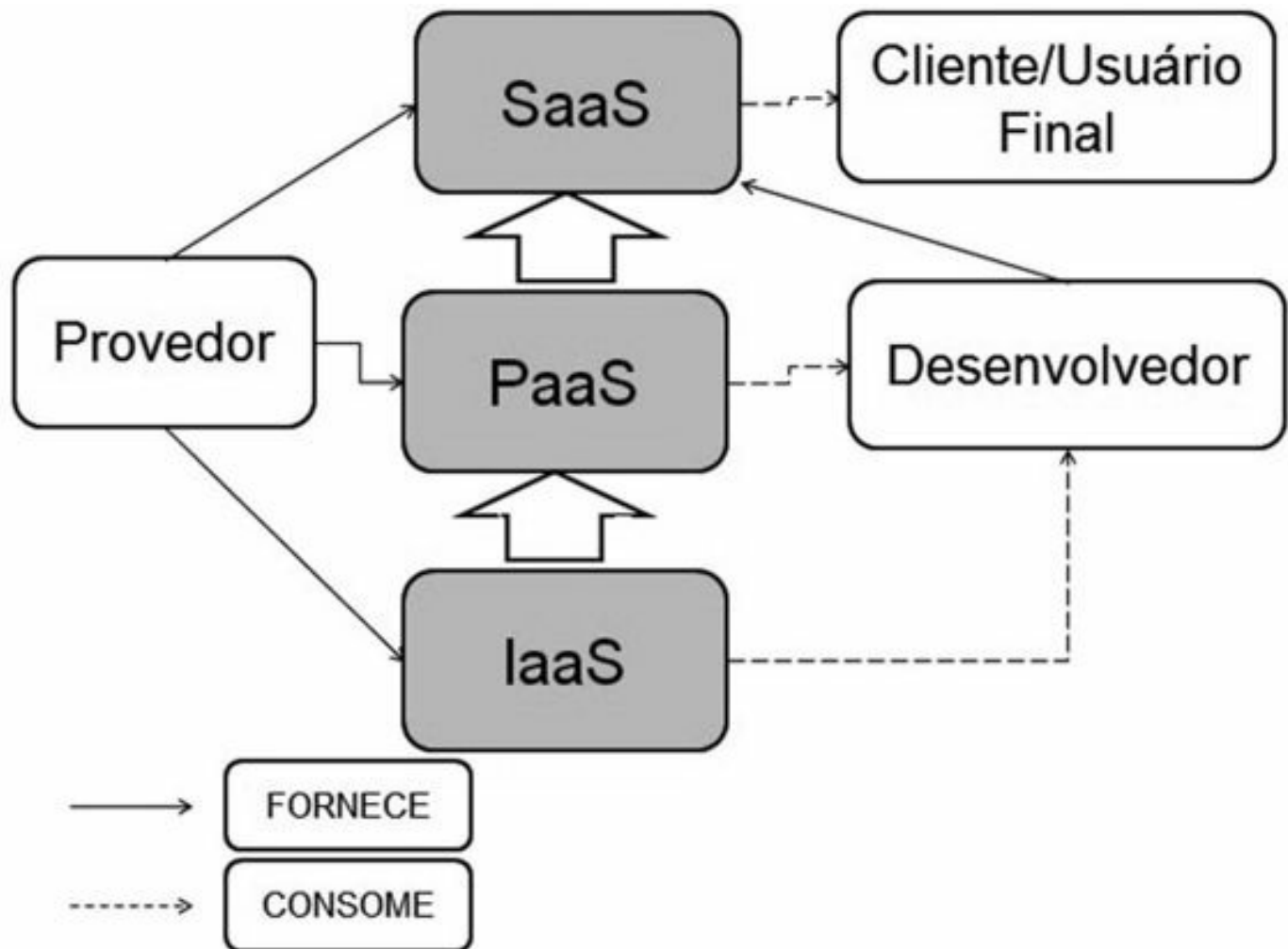


Figura 2-8 – Papéis em CLOUD COMPUTING

O provedor de serviços ideal é responsável por disponibilizar, gerenciar e monitorar toda a estrutura para a solução de CLOUD COMPUTING, deixando os desenvolvedores e usuários finais sem esses tipos de responsabilidade. Para isso, o provedor pode fornecer serviços nas três modalidades de serviços. Os desenvolvedores utilizam os recursos fornecidos e provêm serviços para os usuários finais. Os clientes pagam a conta.

Esta organização em papéis, ilustrada na **Figura 2-8**, ajuda a definir os atores e os seus diferentes interesses em CLOUD COMPUTING. Os atores podem assumir vários papéis ao mesmo tempo de acordo com os interesses, sendo que apenas o provedor fornece suporte a todos os modelos de serviços.

Do ponto de vista de interação entre os três modelos de serviços, a IaaS fornece recursos computacionais, seja de hardware ou software, para a PaaS, que por sua vez fornece recursos, tecnologias e ferramentas para desenvolvimento e execução dos serviços implementados a serem disponibilizados como SaaS. Importante ressaltar que uma organização provedora de serviços de nuvem não precisa obrigatoriamente disponibilizar os três modelos. Ou seja, um provedor de serviços de nuvem pode disponibilizar a opção IaaS sem ter que obrigatoriamente disponibilizar também uma plataforma PaaS.

Uma forma de visualizar os modelos de serviço é entender a pilha de componentes envolvidos com os principais serviços de nuvem e comparar com a opção interna (*on-premises*).

A **Figura 2-9** ilustra os componentes dos serviços de CLOUD COMPUTING. Importante esclarecer o que são *middleware* e *runtime*. *Runtime* é o software responsável pela execução dos programas e o *middleware* é a designação genérica utilizada para se referir aos softwares que são executados entre as aplicações e os sistemas operacionais. O objetivo do *middleware* é facilitar o desenvolvimento de aplicações, tipicamente as distribuídas, assim como facilitar a integração de sistemas legados ou desenvolvidos de forma não integrada.

À medida que o serviço muda da esquerda para a direita, como ilustra a **Figura 2-9**, o responsável pelo gerenciamento e a segurança vai sendo alterado também. No modelo SaaS a responsabilidade do gerenciamento de todo o *stack* de TI é do provedor e, na verdade, o cliente contrata a segurança necessária do provedor, que é o responsável pelo serviço de segurança. No modelo IaaS o cliente é responsável pela implantação e pelo gerenciamento da segurança.

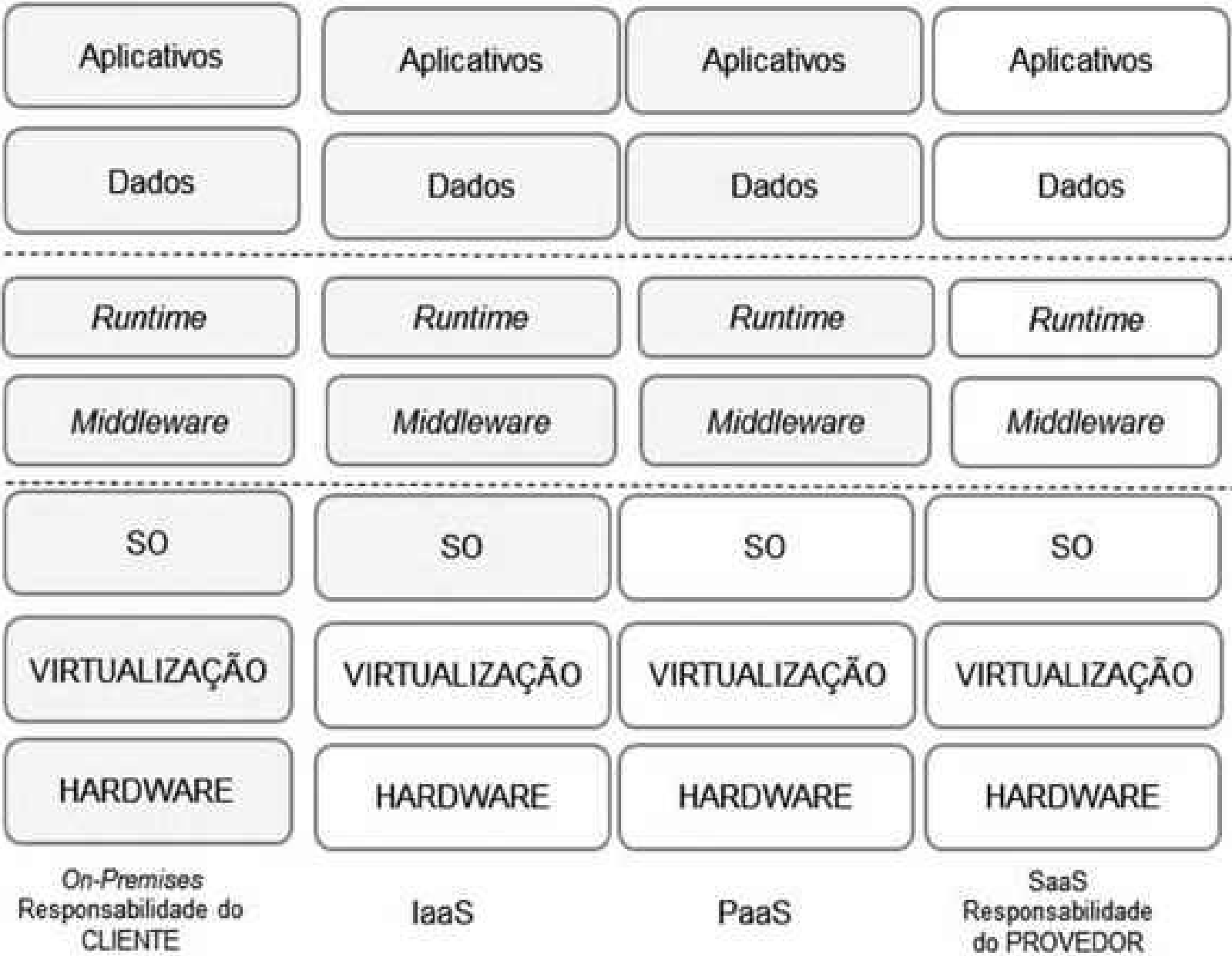


Figura 2-9 – Componentes dos Serviços de CLOUD COMPUTING

Outro aspecto importante é definir onde os serviços de nuvem serão implementados. A **Figura 2-10** ilustra as três opções. Os serviços podem rodar internamente dentro da organização, podem rodar em provedores regionais ou em provedores globais. A definição de onde os serviços de nuvem vão rodar depende de alguns fatores, incluindo aspectos de latência das aplicações e aspectos institucionais. A latência pode obrigar a aplicação a estar mais perto do usuário. As questões institucionais definem muitas vezes onde os dados devem ser armazenados.

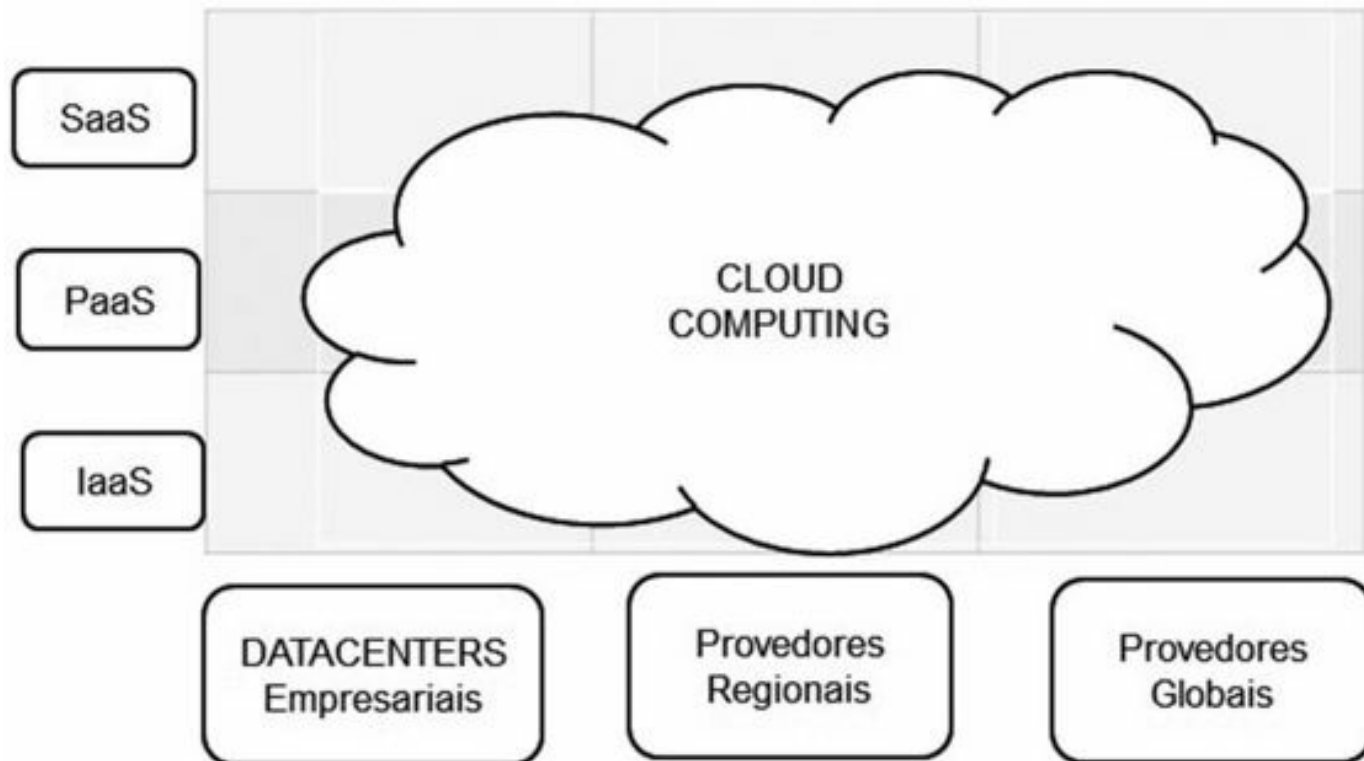


Figura 2-10 – Serviços de CLOUD versus Localização

A Microsoft, por exemplo, funciona como um provedor global e possui atualmente seis DATACENTERS.

2.5. Modelos de Implantação

Existem quatro principais modelos de implantação de CLOUD COMPUTING, descritos a seguir:

- **Nuvem Privada (Private CLOUD):** compreende uma infraestrutura de CLOUD COMPUTING operada e quase sempre gerenciada pela organização cliente. Os serviços são oferecidos para serem utilizados pela própria organização, não estando publicamente disponíveis para uso geral. Em alguns casos pode ser gerenciada por terceiros. Pesquisa realizada pela Forbes Insights⁶ com 235 CIOs e outros executivos de grandes companhias americanas aponta que 52% das organizações estão avaliando a adoção de nuvens privadas. Sugerem-se dois tipos básicos de nuvem privada:
 - **Nuvem Privada, hospedada pela empresa.** É um modelo interessante quando aspectos de *compliance* e controle precisam ser considerados.
 - **Nuvem Privada, hospedada em provedor de serviço.** É um modelo interessante para aplicações de forma geral e aplicações de missão crítica.
- **Nuvem Pública (Public CLOUD):** é disponibilizada publicamente através do modelo *pay-as-you-go*. São oferecidas por organizações públicas ou por grandes grupos industriais que possuem grande capacidade de processamento e armazenamento. A mesma pesquisa realizada pela Forbes Insights aponta que 38% das organizações estão

avaliando também a adoção de nuvens públicas.

- **Nuvem Comunitária (Community CLOUD):** neste caso, a infraestrutura de CLOUD COMPUTING é compartilhada por diversas organizações e suporta uma comunidade que possui interesses comuns. A nuvem comunitária pode ser administrada pelas organizações que fazem parte da comunidade ou por terceiros e pode existir tanto fora como dentro das organizações.
- **Nuvem Híbrida (Hybrid CLOUD):** a infraestrutura é uma composição de duas ou mais nuvens (privadas, públicas ou comunitárias) que continuam a ser entidades únicas, porém, conectadas através de tecnologias proprietárias ou padronizadas que propiciam a portabilidade de dados e aplicações. A nuvem híbrida impõe uma coordenação adicional a ser realizada para uso das nuvens privadas e públicas.

A Oracle sugere uma comparação entre os modelos de nuvem pública e privada do ponto de vista dos benefícios obtidos:

Benefícios comuns a nuvem públicas e privadas incluiriam:

- Alta Eficiência.
- Alta Disponibilidade.
- Elasticidade.
- Rápida Implementação.

Alguns benefícios exclusivos da nuvem pública incluiriam:

- Custos Iniciais Baixos.
- Economia de Escala.
- Simplicidade para Gerenciamento.
- Pagamento como Despesas Operacionais.

Alguns benefícios da nuvem privada incluiriam:

- Maior controle de segurança, *compliance* e qualidade de serviço.
- Mais fácil integração.
- Custos totais mais baixos.
- Despesas de capital (depreciação incluída) e despesas operacionais.

O IDC⁷ também fez uma pesquisa com gerentes de TI focada em saber deles quais os benefícios que poderiam ser obtidos com a migração para nuvem privada e para nuvem pública. Para a nuvem pública foram apontados redução do TCO, recuperação de desastres e melhor disponibilidade. Para a nuvem privada foram apontados os seguintes benefícios: melhorar a disponibilidade, a recuperação de desastres e o uso dos recursos de TI.

A **Figura 2-11** resume os conceitos referentes a CLOUD COMPUTING consolidados pelo NIST.

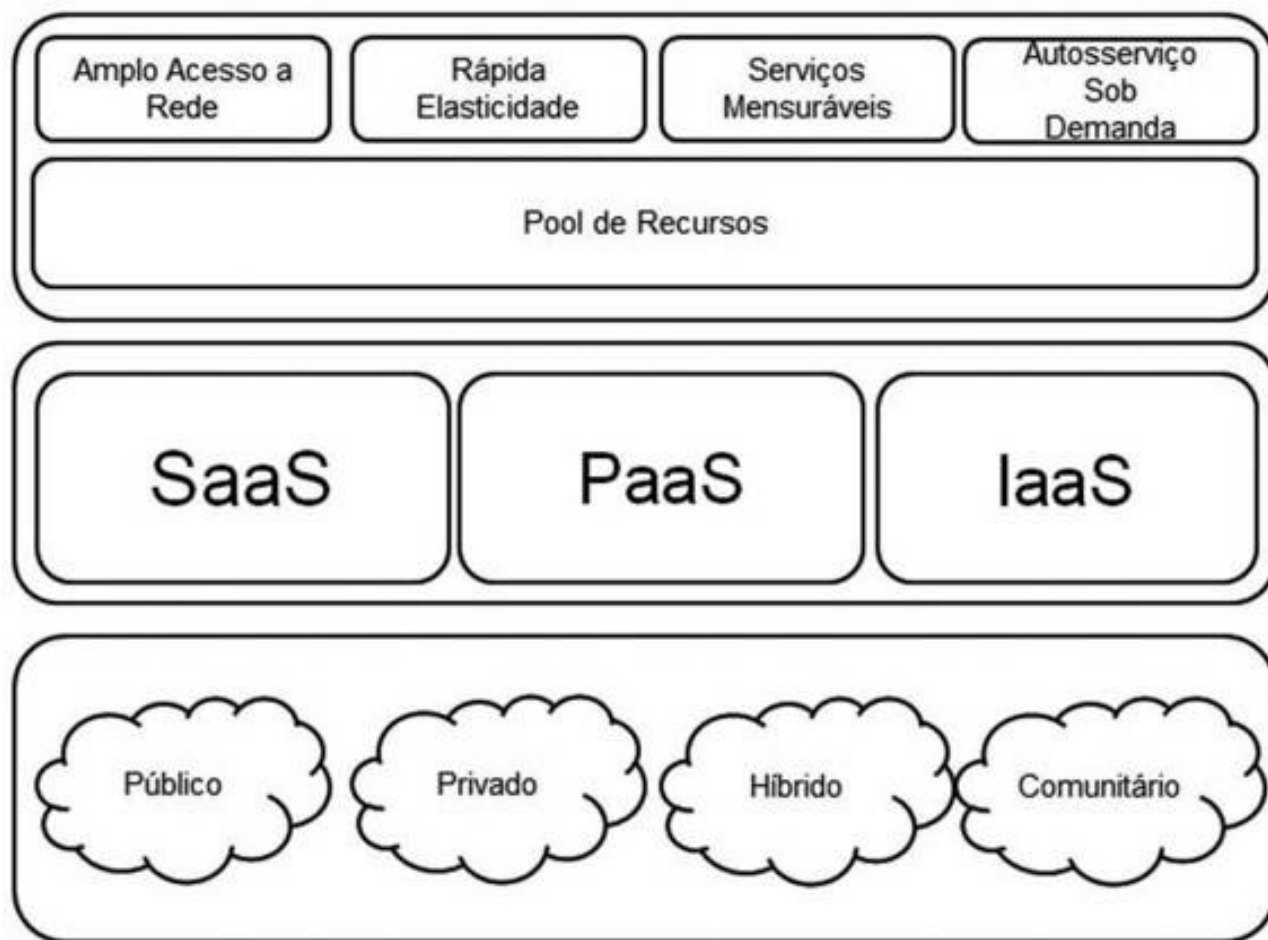


Figura 2-11 – Definição do NIST para CLOUD COMPUTING

2.6. Conceito na Prática: Modelo para Segurança da CSA

O modelo de referência para segurança desenvolvido pela CSA e na versão 2.1 é mostrado na **Figura 2-12**. Este modelo demonstra de forma abrangente as relações e dependências entre os modelos de serviços de CLOUD COMPUTING. O modelo é fundamental para compreender os riscos de segurança envolvidos em uma solução baseada em CLOUD COMPUTING.

O modelo serve para entender os riscos envolvidos com a CLOUD COMPUTING, mesmo que a forma que alguns provedores implementem os serviços não seja exatamente assim.

O modelo IaaS inclui os recursos de infraestrutura desde as instalações até as plataformas de hardware que ali residem. Incorpora a capacidade de abstrair recursos e oferecer conectividade física e lógica a estes recursos. Fornece também um conjunto de APIs que permitem a gestão e outras formas de interação com a infraestrutura por parte dos clientes.

O modelo PaaS acrescenta uma camada adicional de integração com frameworks de desenvolvimento de aplicativos, recursos de *middleware* e funções como banco de dados, mensagens e filas, permitindo aos desenvolvedores criarem aplicativos para a plataforma cujas linguagens de programação e ferramentas são suportadas pela pilha.

O modelo SaaS fornece um ambiente operacional autocontido usado para entregar todos os recursos do usuário, incluindo o conteúdo, a apresentação, as aplicações e a capacidade de gestão.

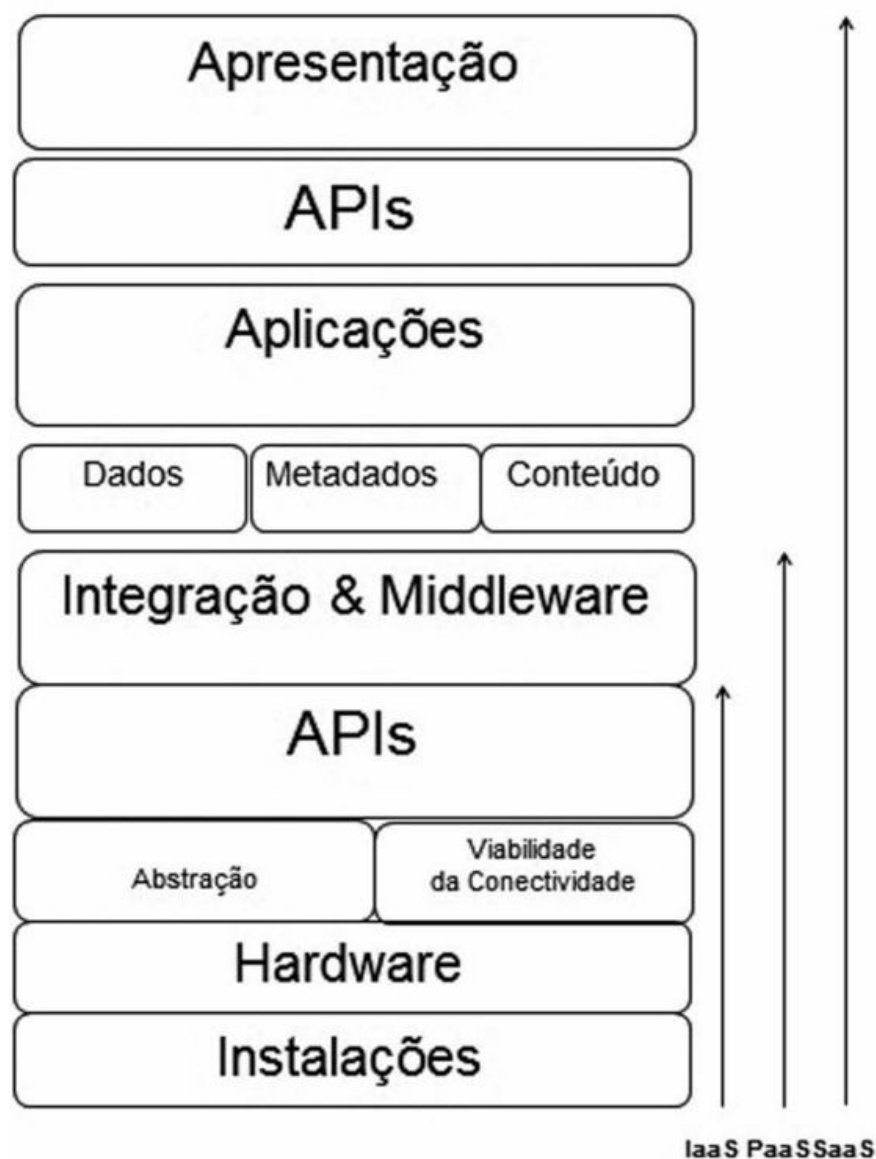


Figura 2-12 – Modelo de Referência para Segurança de CLOUD
(Fonte: GUIA CSA versão 2.1. Adaptação feita por Manoel Veras)

2.7. Arquitetura Multitenancy ou Multi-Inquilino

O guia CSA, citado anteriormente, *reforça* a necessidade de definir o que seja arquitetura *multitenancy* ou multi-inquilino.

Arquitetura *multitenancy* trata da arquitetura de aplicações onde uma única *instance* (instância) do software roda em um servidor e vários tenants (inquilinos) a acessam. Diferentemente da VIRTUALIZAÇÃO, na arquitetura *multitenancy* os inquilinos utilizam a mesma instância do servidor e não máquinas virtuais distintas. No caso da nuvem, a arquitetura *multitenancy* implica em poder forçar a aplicação de políticas, segmentação, isolamento, governança, níveis de serviço de forma diferente por perfis de usuários. As preocupações com segurança *multitenancy* servem tanto para provedores de nuvem pública como para nuvens privadas que possuem unidades de negócio que

precisam ser separadas logicamente, mas que utilizam a mesma infraestrutura.

A arquitetura *multitenancy* sugere que o provedor de nuvem tenha uma abordagem de design e arquitetura que permita economia de escala, disponibilidade, segurança, isolamento, eficiência operacional, aproveitando o compartilhamento da infraestrutura, dos dados e serviços através de diferentes clientes.

A qualidade do serviço SaaS depende da arquitetura *multitenancy*. A proposta do modelo de serviço SaaS é ter uma aplicação atendendo a múltiplos *tenants* ou inquilinos. Importante ressaltar que inquilinos não são usuários individuais, mas empresas clientes do software.

2.8. Iniciativas

Diversas iniciativas de CLOUD COMPUTING surgem. Trata-se aqui das principais iniciativas mundiais nesta área.

- Google e Salesforce são provedores globais que estão focados em provimento de software como serviço (SaaS) e plataforma como serviço (PaaS).
- Amazon é principalmente fornecedor de Infraestrutura como Serviço (IaaS).
- VMware é fornecedora de produtos de infraestrutura para DATACENTERS empresariais e provedores regionais. Que entregam IaaS.
- Microsoft possui a oferta mais completa, funcionando como provedor global para soluções SaaS e PaaS. Mas também entregando soluções para DATACENTERS e provedores regionais.

A **Figura 2-13** ilustra o posicionamento dos principais atores tecnológicos envolvidos em fornecer serviços de nuvem:

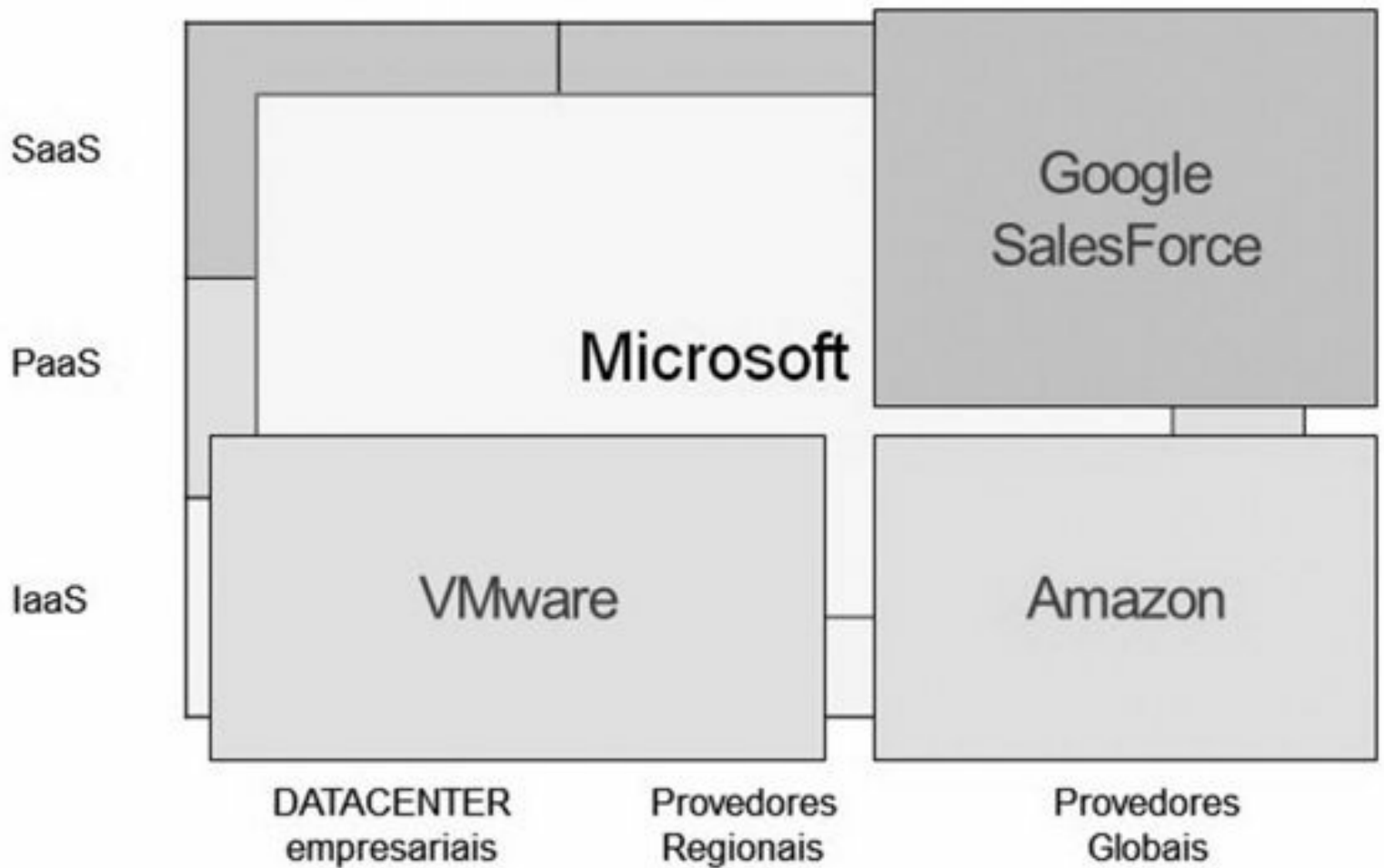


Figura 2-13 – Iniciativas dos Fornecedores de Nuvem

2.9. Referências Bibliográficas

Amazon Web Services disponível em www.amazon.com/aws.

CLOUD COMPUTING: Business Benefits with Security, Governance and Assurance Perspectives. ISACA, an ISACA Emerging Technology White Paper, 2009.

Flávio R. C. Sousa, Leonardo O. Moreira e Javam C. Machado. Computação em Nuvem: Conceitos, Tecnologias, Aplicações e Desafios, 2009.

Greenberg, Albert; Hamilton, James; Maltz, David A.; Patel, Parveen. The Cost of Cloud: Research problems in DATACENTER Networks, 2009.

Saten, John. Deliver CLOUD Benefits Inside your Walls, Forrester, 2009.

The NIST Definition of CLOUD COMPUTING (Draft), NIST, January 2011.

UC Berkeley Reliable Adaptive Distributed Systems Laboratory. Above the Clouds: A Berkeley View of CLOUD COMPUTING, US Berkeley Reliable Adaptative Distributed Systems Laboratory, February 2009.

Vaquero, Luis M & Rodero-Merino, Luis & Caceres, Juan Lindner, Maik. A break in the Clouds: Towards a CLOUD Definition. ACM Sigcomm Computer Communication Review, Volume 39, number 1, January 2009.

- Varia, Jinesh. Architecting for the Cloud: Best Practices, Amazon Web Services, January 2010.
- Veras, Manoel. DATACENTER: Componente Central da infraestrutura de TI, Brasport, 2009.
- Veras, Manoel. VIRTUALIZAÇÃO: Componente Central do DATACENTER, Brasport, 2011.
- Verdi, Fábio Luciano; Rothenberg, Christian; Pasquini, Rafael, Magalhães, Maurício Ferreira, Novas
Arquiteturas de DATACENTER para CLOUD COMPUTING, 2010.

3. Benefícios e Riscos

3.1. Introdução

Este capítulo trata dos benefícios e riscos da CLOUD COMPUTING.

O principal benefício da CLOUD COMPUTING é o ganho de escala propiciado pela arquitetura. Por exemplo, servidores sem uso representam um problema tanto no aspecto do gerenciamento quanto no aspecto do consumo de energia. Servidores a plena carga e servidores a baixa carga consomem energia de forma próxima⁸, portanto, servidor sem uso é sinônimo de ineficiência. Na nuvem, a utilização destes servidores seria otimizada.

3.2. O Benefício da Economia de Escala

Uma forma de entender a economia de escala propiciada pela CLOUD COMPUTING é entender os três principais aspectos envolvidos:

- Economia de escala do lado do fornecimento.
- Economia de escala do lado da demanda.
- Economia de escala da arquitetura multitenancy.

3.2.1. Economia de Escala do Lado do Fornecimento

A economia de escala do lado do fornecimento é propiciada por grandes DATACENTERS que baixam o custo por servidor.

Ocorre principalmente devido a:

- **Redução do custo de energia:** grandes provedores podem localizar seus DATACENTERS em lugares convenientes em termos de melhorar a eficiência energética e taxas de eletricidade, reduzindo o custo quando comparado às opções de manter DATACENTERS com pouca utilização média em grandes companhias.
- **Redução dos custos de pessoal:** devido à escala conseguida com CLOUD COMPUTING, acaba por ocorrer uma redução nos custos de pessoal. Um administrador de CLOUD COMPUTING pode gerenciar mais servidores do que um administrador de um único DATACENTER.
- **Aumento da segurança e confiabilidade:** aqui novamente o ganho de escala possibilita obter uma infraestrutura mais confiável em termos de alta disponibilidade e segurança

física e lógica. Provedores de nuvem deveriam ter, pelo menos em tese, mais *expertise* para resolver problemas de segurança do que um pequeno departamento de TI.

- **Aumento do poder de compra:** operadores de grandes DATACENTERS, componentes da CLOUD COMPUTING, podem utilizar o poder de barganha para adquirir componentes de TI em uma base melhor de preço. Caso contrário, estão fazendo suas organizações perder dinheiro.

3.2.2. Economia de Escala do Lado da Demanda

A economia de escala do lado da demanda ocorre principalmente devido à agregação de demanda que atenua o problema da variabilidade de carga e permite aumentar a taxa de uso do servidor.

Em resumo, a economia de escala do lado da demanda ocorre devido a:

- **Redução da variabilidade:** no lado da demanda, o uso da capacidade é o fator determinante da economia de escala. A VIRTUALIZAÇÃO aperfeiçoa o uso dos recursos e propicia economia de escala principalmente quando as cargas de trabalho variam através do tempo. Mas em A arquitetura CLOUD COMPUTING outras possibilidades otimizam ainda mais o uso da capacidade do sistema, como a variabilidade de uso por parte dos usuários de certos aplicativos como o correio eletrônico. Para obedecer aos níveis de serviço muitas vezes deve-se considerar em instalações corporativas a chance de usuários utilizarem tarefas ao mesmo tempo, forçando o provisionamento de recursos para cima. Em CLOUD COMPUTING o pool de servidores poderia ser mais bem alocado considerando que os recursos são globais.
- **Padrões de uso durante o dia:** os padrões de uso da nuvem em países com fusos horários diferentes são diferentes. Pode-se melhor aproveitar os recursos considerando que horários de pico no Japão são diferentes do horário de pico no Reino Unido para cargas de trabalho semelhantes.
- **Variabilidade da indústria:** indústrias possuem um padrão de uso de aplicativos diferente. A relação entre o uso de recursos de TI no pico e na média pode chegar a 10x na indústria de serviços e ser de 4x no varejo. Ou seja, indústrias utilizam a TI de forma diferente. CLOUD COMPUTING poderia otimizar o uso dos recursos utilizados por indústrias diferentes.
- **Incerteza no crescimento do uso:** sempre se faz um provisionamento para cima em função de não se saber exatamente como cresce a demanda dos usuários para determinado aplicativo. O resultado é que sempre sobram recursos, mesmo em uma escala menor com a VIRTUALIZAÇÃO. CLOUD COMPUTING pode melhorar ainda mais este aspecto da computação.
- **Variabilidade multirrecurso:** a utilização dos recursos tecnológicos como memória e processamento e I/O de servidores varia de acordo com a demanda dos aplicativos. Em um pool de recursos esta característica poderia ser otimizada para melhorar o uso.

3.2.3. Economia de Escala da Arquitetura Multitenancy

O modelo de aplicação *multitenancy* ou multi-inquilino aumenta o número de inquilinos por aplicativo, reduzindo a necessidade de gerenciamento e o custo do servidor por inquilino.

As economias de escala sugeridas nos dois itens anteriores independem da arquitetura do aplicativo, que pode ser do tipo *scale-up* ou *scale-out* ou *multitenant*. Se o aplicativo tiver sido escrito para ser *multitenant* existe aqui outra economia de escala. Os dois principais benefícios são:

- Custo do aplicativo amortizado, por ser compartilhado por vários clientes. O custo de aplicação de *patches*, por exemplo, pode ser reduzido.
- Custo de utilização de servidores amortizado e compartilhado por vários clientes.

3.3. Outros Benefícios

Os outros possíveis benefícios do uso de um modelo de CLOUD COMPUTING são:

- **Aumento da segurança:** uma infraestrutura centralizada pode melhorar a segurança, incluindo rotinas de backup otimizadas e testadas. Há controvérsias.
- **Acesso a aplicativos sofisticados:** aplicativos considerados caros podem ser utilizados no modelo sob demanda.
- **Aumento da produtividade por usuário:** usuários podem acessar os aplicativos de qualquer lugar, o que pode impactar positivamente a produtividade.
- **Aumento da confiabilidade:** existência de estrutura de contingência quase que obrigatória em CLOUD COMPUTING pode melhorar a confiabilidade dos aplicativos.

3.4. Riscos

Riscos de CLOUD COMPUTING é a possibilidade que algum evento imprevisto, falha, ou mesmo mau uso, ameace um objetivo de negócio.

Projetos de adoção de CLOUD COMPUTING, como qualquer projeto, apresentam características conflitantes. Materializar uma melhoria e o risco inerente à consecução desta melhoria são atributos inseparáveis. Assim, os objetivos de um projeto de adoção de CLOUD COMPUTING devem ser acompanhados do gerenciamento dos riscos associados.

Existem várias características importantes no modelo de CLOUD COMPUTING e a combinação destas características fornecidas pelo modelo adotado pelos provedores faz o risco variar.

O Gartner recentemente sugeriu alguns cuidados que o cliente deve ter para mitigar o risco referente à aquisição de serviços de um provedor de CLOUD COMPUTING, descritos a seguir:

- Saber como é feito o acesso dos usuários.
- Saber como o provedor obedece às normas de regulação.

- Saber onde se localizam os dados.
- Saber como os dados são segregados.
- Saber como os dados são recuperados.
- Saber como é feito o suporte.
- Entender a viabilidade do provedor no longo prazo.

O aspecto chave a ser avaliado na opção de entregar os serviços de TI para um fornecedor de CLOUD COMPUTING é o risco da perda do controle dos dados internos, e ele de fato existe.

Mesmo aplicativos que eventualmente continuem a rodar localmente podem utilizar serviços de infraestrutura providos pela nuvem, como armazenamento de dados, por exemplo, e funcionar utilizando serviços providos internamente e externamente. Normalmente os riscos aumentam até nestes casos.

O que deve ficar claro é que o ambiente de CLOUD COMPUTING é essencialmente diferente do ambiente tradicional de computação. Muda-se de um modelo amparado em equipamentos para um modelo orientado a serviços. Os acordos de nível de serviço (*Service Level Agreement – SLA*) passam a ser a interface natural entre provedores e organizações clientes. Os SLAs podem ser definidos para a aplicação ou para a infraestrutura de TI.

A computação baseada em nuvem pública adiciona novos atores à equação de segurança – provedores de serviço terceirizados, fornecedores de infraestrutura e funcionários diminuindo o controle que a TI exerce sobre estas três áreas.

As ofertas de serviços de nuvem ainda são confusas e, na medida em que estas ofertas crescem, os aspectos de segurança precisam ser rapidamente considerados. A oferta na forma de nuvem é uma opção ao DATACENTER corporativo e é preciso considerar os novos atores nesta nova forma de adquirir serviços de TI. A confiança é o aspecto central na ida para a nuvem.

Os serviços de CLOUD COMPUTING precisam ser elásticos conforme as necessidades dos clientes, precisam ser adequados à realidade dos clientes e às exigências das suas aplicações e precisam ser oferecidos em diversos locais. Tudo isto deve ser oferecido com segurança. A **Figura 3-1** ilustra estas características.

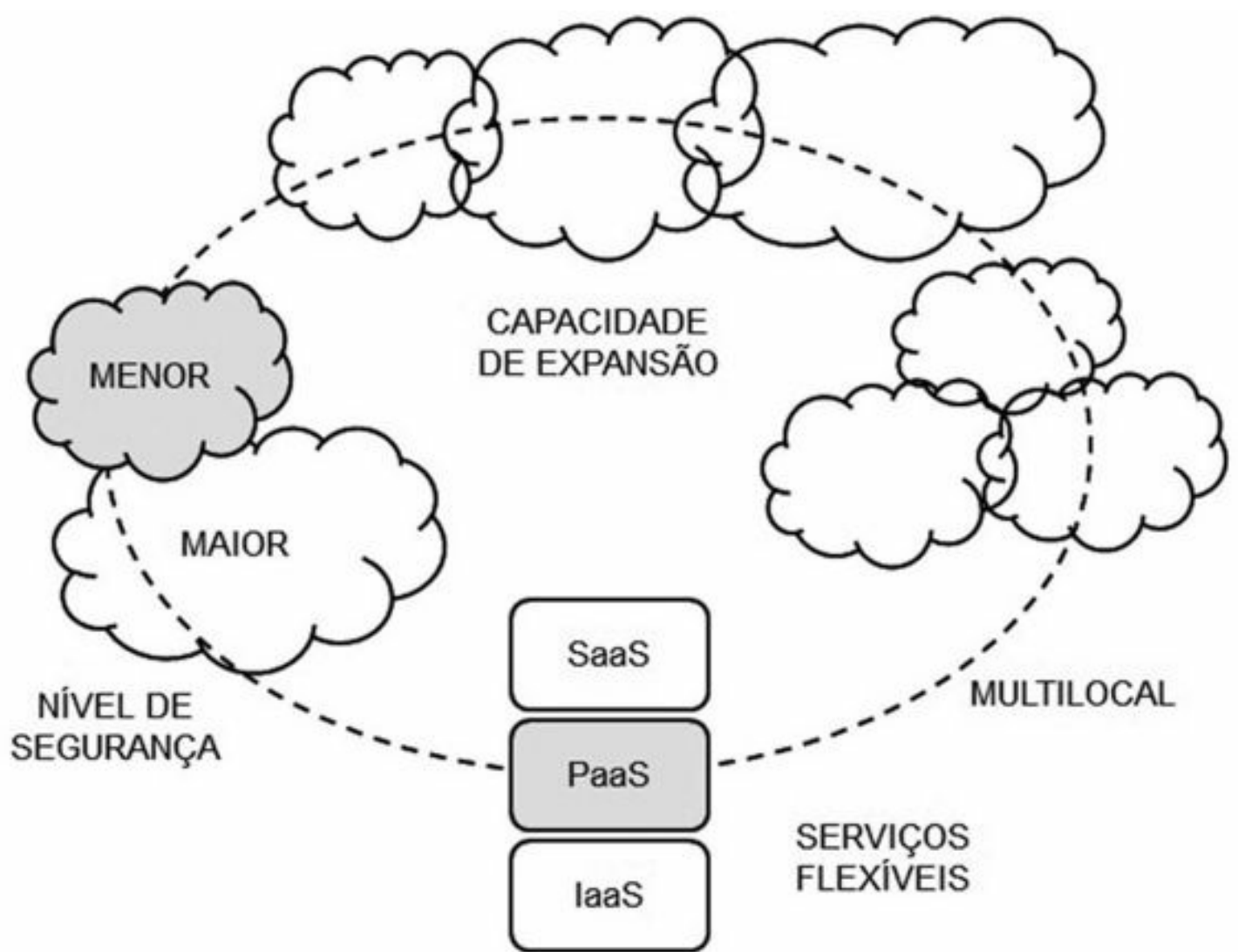


Figura 3-1 – Recursos Essenciais de CLOUD (Fonte: RSA)

Os aplicativos que estão indo para a nuvem ainda são de pouca confidencialidade, mas, com o aumento exponencial da base de informação digital, as empresas precisam considerar o fato de que respondem por boa parte destas informações em termos de segurança, privacidade e confidencialidade. O crescimento de uso da nuvem só será possível com o aumento da segurança. As empresas precisam assegurar a confidencialidade, a integridade e a disponibilidade dos dados no momento em que eles forem transmitidos, armazenados ou processados por terceiros na cadeia de serviços em nuvem.

Aspectos de segurança da infraestrutura tradicional mudam com a adoção da nuvem. As relações de confiança na cadeia da nuvem precisam ser cuidadosamente consideradas. O DATACENTER na nuvem agora é controlado por terceiros e estruturas convencionais de segurança já não servem mais. O controle passa a ser compartilhado e questões de responsabilidade vêm à tona: como saber quem teve acesso a quais informações e aplicações? O acesso precisa ser individualizado. Diversas organizações trabalham mutuamente para definir os padrões de segurança na nuvem. A abrangência vai desde autenticação, autorização delegada, gerenciamento de chaves públicas, proteções contra perdas de dados e emissão de relatórios normativos.

A portabilidade na nuvem é outra questão ainda a ser resolvida. Serão necessários arranjos que permitam portabilidade de identidade e informações entre várias nuvens. As questões de

confidencialidade e privacidade são também muito relevantes. Abordagens de controle de acesso usuais em operações convencionais podem não funcionar na nuvem. Controle de acesso baseado em conteúdo e não na identidade do usuário é uma questão ainda não bem resolvida. O controle sobre informações confidenciais é uma questão muito relevante também. As empresas precisam saber se os provedores de serviço atendem à sua política de conformidade. Correlação de eventos internos e externos é outra demanda. Quem é o responsável pela violação dos dados?

Os níveis de serviço de segurança dos dados também deverão variar e devem atender a diferentes necessidades de confidencialidade otimizando o TCO. A segurança deverá ser mantida durante todo o ciclo de vida dos dados.

Os elementos principais para a proteção na nuvem são:

- **Segurança de identidade:** o gerenciamento completo de identidades, os serviços de autenticação de terceiros e a identidade federada se tornarão elementos fundamentais para a segurança da nuvem. A segurança da identidade preserva a integridade e a confidencialidade dos dados e dos aplicativos enquanto deixa o acesso prontamente disponível para os usuários apropriados. O suporte a esses recursos de gerenciamento de identidade para usuários e componentes da infraestrutura será um dos principais requisitos da computação em CLOUD COMPUTING e a identidade precisará ser gerenciada de maneira que gere confiança.
- **Segurança das informações:** no DATACENTER tradicional, os controles sobre o acesso físico, o acesso a hardware e software e os controles de identidade se combinam para proteger os dados. Na nuvem, a barreira protetora que protege a infraestrutura é diluída. Para compensar, a segurança passará a ser centrada nas informações. Os dados precisam de segurança própria que os acompanhe e os proteja.
- **Segurança da infraestrutura:** a infraestrutura de base da nuvem deve ser inerentemente segura, independentemente da nuvem ser privada ou pública ou prover serviço SaaS, PaaS ou IaaS.

Outra forma de ver a segurança é utilizar o modelo de referência proposto anteriormente pela guia CSA versão 2.1. O documento foi elaborado pela CLOUD Security Alliance, organização norte-americana sem fins lucrativos, com o objetivo de orientar sobre o uso de melhores práticas da prestação de serviços na área de segurança na nuvem.

Utilizando-se o modelo da CSA podem ser feitas considerações importantes sobre os três modelos de serviços e a segurança:

- O modelo SaaS oferece funcionalidade mais integrada, construída diretamente baseada na oferta, com a menor extensibilidade para o cliente e um nível relativamente elevado de segurança (fornecedor assume a responsabilidade pela segurança).
- O modelo PaaS visa permitir que os desenvolvedores criem seus próprios aplicativos em cima da plataforma. Assim, é mais extensível que o modelo SaaS, à custa das funcionalidades disponibilizadas aos clientes. As capacidades de segurança são menos completas, mas há flexibilidade para adicionar uma camada de segurança extra.
- O modelo IaaS oferece pouca ou nenhuma característica típica de um aplicativo, mas

permite muita extensibilidade. Significa ter menos recursos e funcionalidades integradas de segurança. Este modelo requer que o cliente proteja e gerencie os sistemas operacionais, aplicativos e conteúdo.

O guia CSA versão 2.1 chama a atenção e endossa o que foi anteriormente descrito: que as modalidades de implantação e consumo da nuvem devem ser pensadas não só no contexto do interno versus externo, nem só considerando a localização física dos ativos, recursos e informações, mas também no contexto de quem são os clientes e de quem é responsável por questões de governança, segurança e conformidade com políticas e padrões.

Aspecto de localização dentro ou fora da empresa de um ativo é relevante para a segurança, mas o que o guia citado chama a atenção é que os riscos dependem também de:

- Tipos de ativos, recursos e informações que estão sendo gerenciados;
- Quem as gerencia e como as gerencia;
- Quais controles estão selecionados e como estão integrados;
- Questões de conformidade.

A **Figura 3-2** sugere que o mapeamento de serviço de computação de nuvem pode ser comparado a um modelo de controle de segurança e a um modelo de conformidade para determinar quais controles existem e quais não existem. Pode-se assim comparar a arquitetura de segurança existente, bem como os requisitos de negócio, de regulamentação e outros requisitos de conformidade com o padrão. Tipicamente um exercício de *gap analysis*.

Uma vez a *gap analysis* realizada, baseada em requisitos de controle ou exigência de conformidade, torna-se mais fácil determinar o que precisa ser feito em termos de risco: aceitação, transferência ou mitigação. Define-se um plano de melhorias e parte-se para a execução.

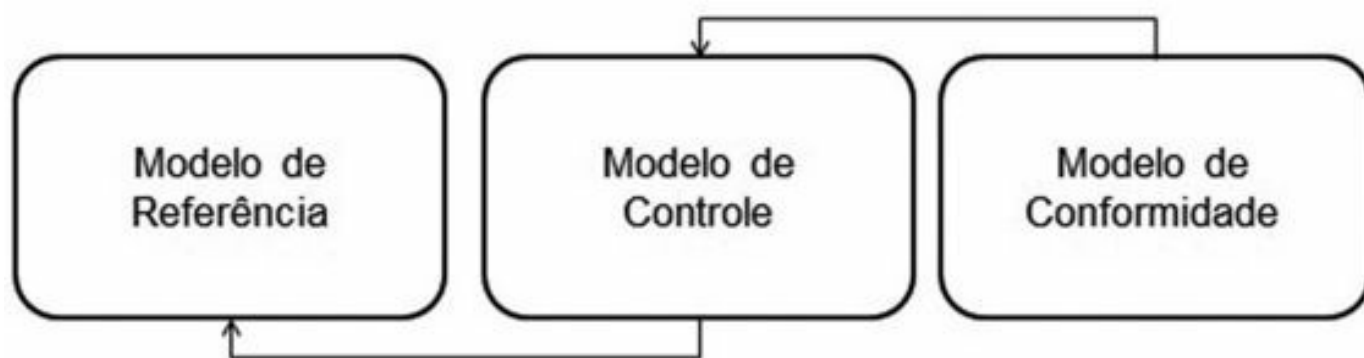


Figura 3-2 – Modelos de Referência, Segurança e Conformidade (Fonte GUIA CSA versão 2.1)

Importante ressaltar que as responsabilidades de segurança do provedor e do cliente diferem muito de um modelo para outro. No caso da Amazon EC2, um típico caso de IaaS, toda segurança relacionada com o sistema de TI, incluindo o Sistema Operacional (SO), aplicativos e dados, corre por conta do cliente. No caso da solução de CRM da Salesforce, fornecido em modelo SaaS, toda a segurança é de responsabilidade do provedor.

A ideia de reutilização, ganhos de escala e padronização do modelo de CLOUD COMPUTING que teoricamente possibilita obter menores custos vai de encontro a certa rigidez, quase sempre imposta pela segurança.

Em SaaS, os controles de segurança e os escopos envolvidos são negociados em contratos baseados em acordo de nível de serviços (SLAs) legalmente tratados. Em IaaS, a responsabilidade de proteger a infraestrutura básica e camadas de abstração pertence ao provedor, mas o restante da pilha é de responsabilidade do consumidor. Já o modelo PaaS oferece um equilíbrio intermediário, onde quem garante a plataforma é o provedor e a segurança das aplicações desenvolvida e a tarefa de desenvolvê-la de forma segura é do cliente.

Para ajudar empresas nessa tarefa, o Guia de Segurança da CSA versão 2.1 traz algumas recomendações:

- É preciso examinar e avaliar a cadeia de suprimentos do fornecedor. Isso também significa verificar o gerenciamento de serviços terceirizados pelo próprio fornecedor.
- A avaliação dos fornecedores de serviços terceirizados deve concentrar-se nas políticas de recuperação de desastres e continuidade de negócio e em processos. Deve incluir também a revisão das avaliações do fornecedor destinadas a cumprir exigências de políticas e procedimentos e a avaliação das métricas usadas pelo fornecedor para disponibilizar informações sobre o desempenho e a efetividade dos controles.
- O plano de recuperação de desastres e continuidade de negócios do cliente deve incluir cenários de perda dos serviços prestados pelo fornecedor e de perda pelo prestador de serviços terceirizados e de capacidades dependentes de terceiros.
- A regulamentação da governança de segurança de informações, a gestão de riscos e as estruturas e os processos do fornecedor devem ser amplamente avaliados.
- É preciso solicitar documentação sobre como as instalações e os serviços do fornecedor são avaliados quanto aos riscos e auditados sob controles de vulnerabilidades. Além disso, deve-se procurar solicitar uma definição do que o fornecedor considera fator de sucesso de segurança da informação e serviços críticos, indicadores-chave de desempenho e como esses pontos são mensurados.

Artigo publicado na Computerworld americana, traduzido para o português com o título “Considere os riscos legais antes de contratar o serviço de CLOUD”, reforça que uma organização não deve assinar contratos de CLOUD COMPUTING sem levar em conta cinco questões fundamentais: privacidade, conformidade jurídica, mandados de busca, guarda de dados sensíveis e segurança da informação. Os principais aspectos relativos a estas questões e tratados no referido artigo são transcritos a seguir:

- **Privacidade:** para se proteger, os clientes precisam conferir se os provedores são capazes de atender às regras a que estão submetidos e insistir que tudo seja descrito no contrato.
- **Conformidade com múltiplas jurisdições:** os clientes da nuvem devem saber a localização do fornecedor e de seus servidores para determinar onde poderiam sofrer com problemas jurídicos. Ninguém quer se ver em meio a uma disputa judicial em outro

estado ou outro país pelo fato de ter descuidado deste aspecto.

- **Mandados de busca:** dados de múltiplos clientes podem acabar no mesmo servidor. Caso um dos clientes sofra um processo que gere um mandado por buscas naquele servidor, todos os dados podem acabar se tornando inacessíveis para a companhia, que não tem nada a ver com aquilo. A empresa precisa ter um plano para minimizar esses riscos e garantias do provedor de que as informações estão particionadas de forma a não afetar os dados dos outros clientes, caso apenas um deles se envolva em uma questão judicial.
- **Informações legais:** o detentor de uma informação tem a obrigação de preservar qualquer dado que possa ser relevante em litígios, mantendo-o armazenado para propósitos legais. Um exemplo são as informações trabalhistas: se não tiver registro de todos os dados de seus funcionários, a empresa pode ser acionada legalmente para prestar informações, caso sofra ação. E a justiça pode ir diretamente ao provedor, com um mandado judicial, de forma que a empresa perde o controle de toda a situação. Além disso, buscar informações pode ser difícil se o provedor não tiver procedimentos muito bem documentados de armazenamento, facilitando a consulta futura. A empresa deve ter a capacidade de buscar os documentos exatos que importam ao pedido, para não sofrer com multas. E o contrato com o fornecedor de nuvem deve ser capaz de prever isso com exatidão.
- **Segurança da informação:** os métodos para proteger dados na nuvem, como criptografia, estão bem documentados. Mas há também riscos associados com a manutenção de todos os registros da empresa em uma dada localização.

Outra questão: quem paga pelos custos associados a brechas de segurança na nuvem? O cliente normalmente espera que o fornecedor pague a conta, mas, em muitos locais, a empresa que armazena dados na nuvem é a responsável por notificar clientes na eventualidade de brechas de segurança. E é por isso que ela precisa que o fornecedor se comprometa, em contrato, a avisar de Ocorrências imediatas de segurança. Deve-se especificar muito bem de quem é a responsabilidade no caso de qualquer brecha de segurança.

3.5 Consolidação dos Riscos

Sidney Chaves (2011), em sua dissertação de mestrado defendida na FEA-USP, fez uma boa consolidação dos riscos inerentes a CLOUD COMPUTING e resumidos a seguir:

- **Riscos operacionais:** falta de privacidade devido, por exemplo, a deficiências de isolamento no ambiente de nuvem, falta de integridade que pode ser provocada por agentes de software, suporte inadequado por parte do provedor ocasionado por uma série de fatores, incluindo pessoal mal preparado, baixo desempenho dos serviços contratados, ataques por saturação não detectados a tempo, dificuldades para provisionar ou liberar recursos e baixa ou nenhuma interoperabilidade entre aplicativos.
- **Riscos de negócio:** indisponibilidade temporária por parte do provedor e até não

continuidade, que seria uma interrupção definitiva por parte do provedor.

- **Riscos estruturais:** não conformidade com padrões e legislação, limitações na forma de realizar o licenciamento de software, aprisionamento feito pelo provedor e má reputação do provedor, oriunda de baixa qualidade de serviços prestados.

3.6. Referências Bibliográficas

Chaves, Sydney. A questão dos riscos em ambientes de computação de nuvem, dissertação de mestrado, FEA-USP, 2011.

CLOUD COMPUTING: Business Benefits with Security, Governance and Assurance Perspectives. ISACA, an ISACA Emerging Technology White Paper, 2009.

Greenberg, Albert; Hamilton, James; Maltz, David A.; Patel, Parveen. The Cost of Cloud: Research problems in DATACENTER Networks, 2009.

O papel da segurança na computação em nuvem confiável. RSA, 2009.

Saad, Alfredo C. Terceirização de Serviços de TI. Brasport, 2006.

Saten John, Deliver CLOUD Benefits Inside your Walls, Forrester, 2009.

The Benefits and Risks of CLOUD Platforms: A Guide for Business Leaders, Microsoft, 2011.

The Economics of the CLOUD. Microsoft, November 2010.

The NIST Definition of CLOUD COMPUTING (Draft), NIST, January 2011.

Varia, Jinesh. Architecting for the Cloud: Best Practices, Amazon Web Services, January 2010.

Veras, Manoel. DATACENTER: Componente Central da infraestrutura de TI, Brasport, 2009.

Veras, Manoel. VIRTUALIZAÇÃO: Componente Central do DATACENTER, Brasport, 2011.

4. Tomada de Decisão

4.1. Introdução

A decisão de utilizar uma arquitetura de nuvem pública não é simples. Estruturar a nuvem privada é a própria evolução do DATACENTER corporativo. Já a decisão de adotar a nuvem pública é muito mais complexa e deve estar aderente ao modelo de governança de TI adotado.

Conforme já dito anteriormente, existem muitos benefícios e riscos associados ao modelo de CLOUD COMPUTING. A ideia central é saber se existe maturidade da organização para utilizar aplicações que utilizam dados armazenados e processados na nuvem.

O modelo a ser utilizado, a forma de implementação e o momento também devem ser cuidadosamente pensados. A governança de TI, seu aspecto formal, ajuda muito a tomada de decisão em relação à adoção de CLOUD COMPUTING, pois pode esclarecer as responsabilidades e direitos decisórios dos envolvidos.

4.2. Governança da TI

A Governança Corporativa é um tema em evidência. Para a OCDE (Organização para a Cooperação e Desenvolvimento Econômico), significa criar uma estrutura que determine os objetivos organizacionais e monitore o desempenho para assegurar a concretização desses objetivos. A governança de TI é um conceito derivado da governança corporativa.

A governança de TI deve refletir a governança corporativa no que diz respeito às necessidades de controle da informação, ao mesmo tempo em que se concentra em amparar a gestão da TI e a gestão dos recursos envolvidos para o atingimento de metas de desempenho e a obediência a normas de regulação. A governança da TI também pode ser vista como um mecanismo para estimular comportamentos desejáveis em TI (WEILL & ROSS, 2010).

O modelo de governança de TI implantado depende do modelo de governança da organização e da estratégia vigente. Ou seja, a estratégia define o modelo de governança a ser utilizado, que, por sua vez, define o modelo de governança de TI. A gestão da TI é um tema mais amplo e a governança de TI uma parte importante da gestão, mas que prioriza fundamentalmente o aspecto do alinhamento entre o negócio e a TI.

Existem diversos modelos de melhores práticas relacionados a temas de governança de TI. O COBIT (*Control Objectives for Information and related Technology*) é o principal deles, sendo uma espécie de ferramenta, um guia, que permite que a organização ganhe tempo, principalmente quando da adoção de um modelo de governança de TI.

O COBIT traz uma série de melhores práticas que o mercado já consagrou e que permitem mais rapidamente a adoção de um modelo de governança. Também para os auditores é uma ferramenta que permite a avaliação do estado da área de TI. O modelo surgiu em 1996 e é organizado por processos. O COBIT foi desenvolvido pelo ISACA (*Information Systems Audit and Control Association*; <http://www.isaca.org/>) e hoje é coordenado pelo *IT Governance Institute* (<http://www.itgovernance.org/>), criado em 1998 e afiliado do ISACA. O COBIT é um conjunto de estruturas e processos que visa garantir que a TI suporte e maximize, adequadamente, os objetivos e estratégias de negócio da organização e está na versão 4.1.

Recentemente uma pesquisa⁹ realizada pela FGV-SP confirmou a aderência das organizações de TI ao COBIT. Segundo a pesquisa, o COBIT (28,3%) é a principal prática de governança da TI utilizada no Brasil, conforme ilustra a **Figura 4-1**.

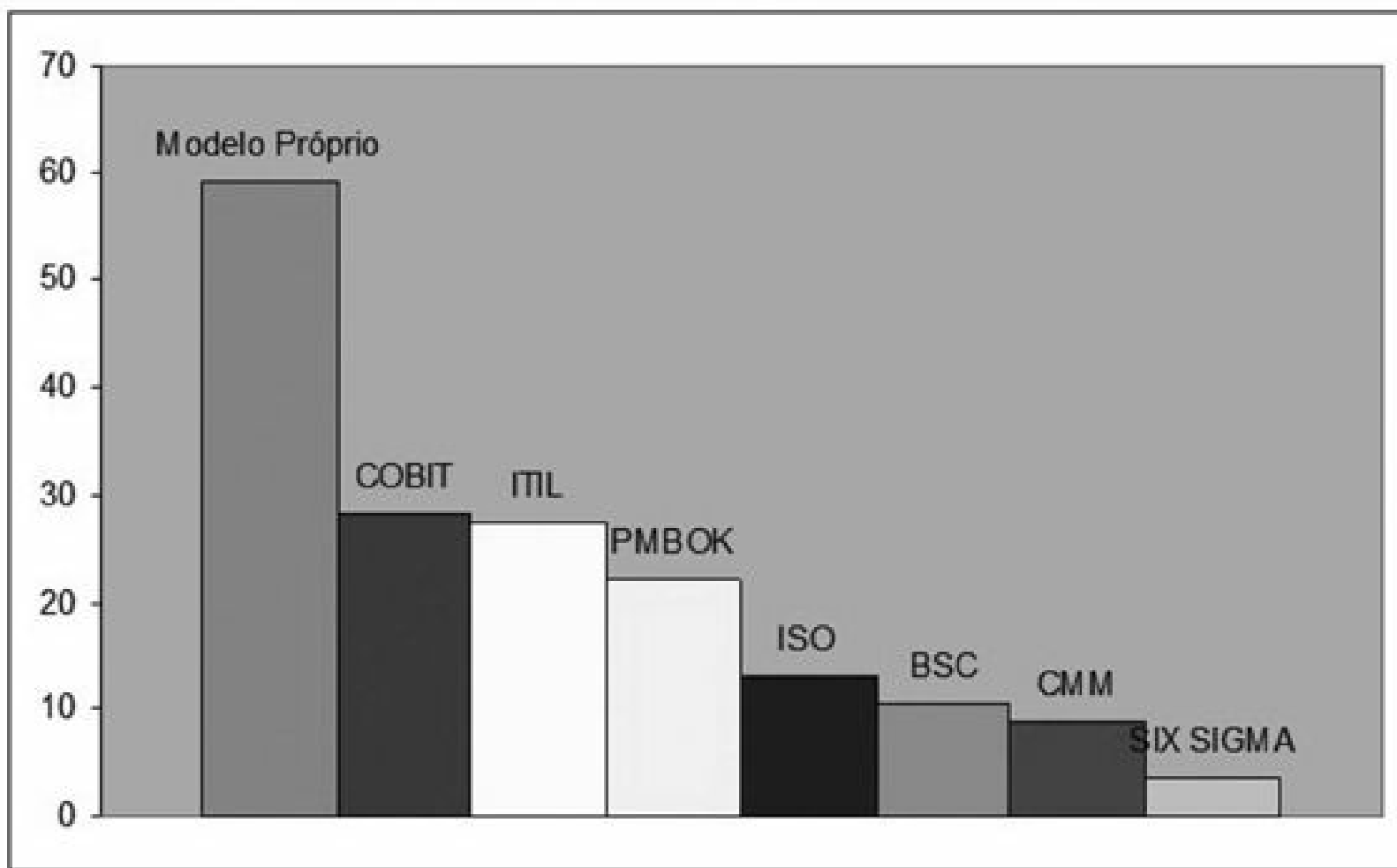


Figura 4-1 – Governança de TI no Brasil

O COBIT define a governança da TI como uma estrutura de relacionamentos entre processos para direcionar e controlar uma empresa de modo a atingir os objetivos corporativos. Para esse gerenciamento e controle o COBIT propõe métodos que utilizam 34 objetivos de controle de alto nível, sendo que, para cada objetivo de controle de alto nível, são definidos vários objetivos de controle detalhados. O método inclui: elementos de medidas de desempenho, fatores críticos de sucesso e modelos de maturidade.

O COBIT foi concebido para ser utilizado em quatro domínios (planejamento e organização,

aquisição e implementação, entrega e suporte, monitoração e avaliação), conforme ilustra a **Figura 4-2**. O guia vincula os requisitos básicos da informação (efetividade, eficiência, confidencialidade, integridade, disponibilidade, fidelidade e confiabilidade) aos recursos de infraestrutura e aplicações de TI. O modelo define uma relação bidirecional entre objetivos do negócio, governança de TI e informação.

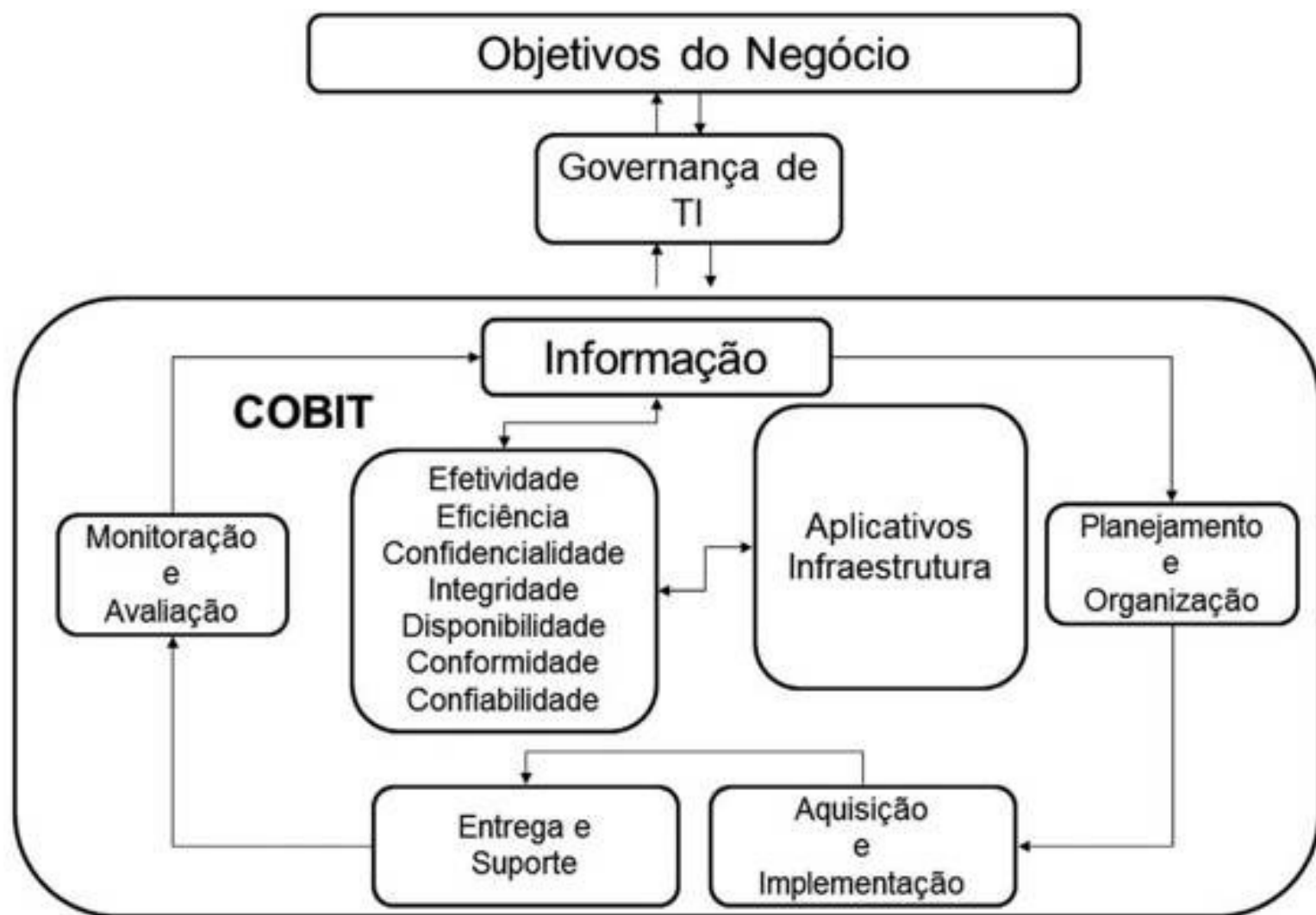


Figura 4-2 – COBIT

Os domínios do COBIT são:

- **Planejamento e Organização:** composto por definir um plano estratégico de TI; definir a arquitetura de informação; determinar a direção tecnológica; definir a organização e relacionamentos de TI; gerenciar o investimento de TI; comunicar os objetivos e direção gerenciais; gerenciar recursos humanos; garantir aderência com requisitos externos; gerenciar riscos; gerenciar projetos; e gerenciar qualidade.
- **Aquisição e Implementação:** composto por identificar soluções automáticas; adquirir e manter aplicações; adquirir e manter infraestrutura de TI; desenvolver e manter procedimentos; implementar e aprovar sistemas; e gerenciar mudanças.
- **Entrega e Suporte:** composto por definir e gerenciar níveis de serviço; gerenciar

serviços terceirizados; gerenciar desempenho e capacidade; garantir continuidade de serviço; garantir segurança de sistemas; identificar e alocar custos; educar e treinar usuários; prover serviço a clientes; gerenciar a configuração; gerenciar problemas e incidentes; gerenciar dados; gerenciar funcionalidades; e gerenciar operações.

- **Monitoração e Avaliação:** composto por monitorar os processos; avaliar a adequação de controle interno; obter garantia independente; e utilizar auditoria independente. Os controles definidos pelo COBIT são políticas, procedimentos, práticas e estrutura organizacional que devem ser seguidos para assegurar que os objetivos de negócio serão alcançados e que eventos indesejáveis serão prevenidos, detectados e corrigidos.

Cada processo pode considerar até sete critérios de informação que são classificados como de impacto primário, secundário ou não relevante. Os critérios de informação são:

- **Efetividade** lida com a informação relevante e pertinente para o processo de negócio, bem como sendo entregue em tempo, de maneira correta, consistente e utilizável.
- **Eficiência** relaciona-se com a entrega da informação através do melhor (mais produtivo e econômico) uso dos recursos.
- **Confidencialidade** está relacionada com a proteção de informações confidenciais para evitar a divulgação indevida.
- **Integridade** relaciona-se com a fidedignidade e totalidade da informação, bem como sua validade de acordo os valores de negócios e expectativas.
- **Disponibilidade** relaciona-se com a disponibilidade da informação quando exigida pelo processo de negócio hoje e no futuro. Também está ligada à salvaguarda dos recursos necessários e capacidades associadas.
- **Conformidade** lida com a aderência a leis, regulamentos e obrigações contratuais aos quais os processos de negócios estão sujeitos, isto é, critérios de negócios impostos externamente e políticas internas.
- **Confiabilidade** relaciona-se com a entrega da informação apropriada para os executivos para administrar a entidade e exercer suas responsabilidades fiduciárias e de governança.

Também são considerados quatro tipos de recursos de TI que podem ou não ser relevantes para cada processo:

- **Informações:** são os dados em todas as suas formas, a entrada, o processamento e a saída fornecida pelo sistema de informação em qualquer formato a ser utilizado pelos negócios.
- **Aplicativos:** são os sistemas automatizados para usuários e os procedimentos manuais que processam as informações.
- **Infraestrutura:** refere-se à tecnologia e aos recursos (ou seja, hardware, sistemas

operacionais, sistemas de gerenciamento de bases de dados, redes, multimídia e os ambientes que abrigam e dão suporte a eles) que possibilitam o processamento dos aplicativos.

- **Pessoas:** são os funcionários requeridos para planejar, organizar, adquirir, implementar, entregar, suportar, monitorar e avaliar os sistemas de informação e serviços. Eles podem ser internos, terceirizados ou contratados, conforme necessário.

O COBIT, para tratar de aspectos de auditoria, utiliza uma sistemática baseada no CMM (*Capability Maturity Model*), modelo de maturidade para desenvolvimento de software proposto pelo SEI (*Software Engineering Institute*), que estabelece níveis para o processo que está sendo auditado.

- **Inexistente:** não implantado;
- **Inicial:** implantado, mas não planejado;
- **Repetível:** implantado mais ainda intuitivo;
- **Definido:** processo realizado, documentado e comunicado para a organização;
- **Gerenciado:** processo monitorado e avaliado;
- **Otimizado:** melhores práticas são utilizadas para melhorar continuamente os processos.

Assim, é possível auditar a governança de TI numa organização, classificar cada processo em termos de maturidade e definir, mediante a escolha de uma prioridade, um plano para melhoria dos processos selecionados. O COBIT permite que a organização ganhe tempo, principalmente quando da adoção de um modelo de governança, pois já traz uma série de melhores práticas que o mercado já consagrou e que permitem mais rapidamente a adoção de um modelo de governança. Também para os auditores é uma ferramenta que permite a avaliação do estado da área de TI. O COBIT também trata de fatores críticos de sucesso, iniciativas que permitem o pleno controle da TI e também faz o vínculo com o BSC (*Balanced Scorecard*) de Kaplan e Norton para a definição de indicadores de desempenho.

A governança de TI deve tratar das seguintes áreas-foco sugeridas pelo COBIT e aqui resumidas:

- **Alinhamento Estratégico:** visa garantir a ligação entre os planos de negócios e de TI, definindo, mantendo e validando a proposta de valor de TI, alinhando as operações de TI com as operações da organização.
- **Entrega de Valor:** visa garantir que a TI entregue os benefícios previstos na estratégia da organização.
- **Gestão de Recursos:** refere-se a melhor utilização dos investimentos e gerenciamento dos recursos de TI, incluindo informações, aplicativos, infraestrutura e pessoas.
- **Gestão de Riscos:** trata de dar transparência aos riscos significantes para a organização e do seu gerenciamento, além de cuidar dos requisitos de conformidade.

- **Mensuração de Desempenho:** trata de acompanhar e monitorar a implementação da estratégia de TI.

A **Figura 4-3** ilustra as áreas-foco do COBIT.

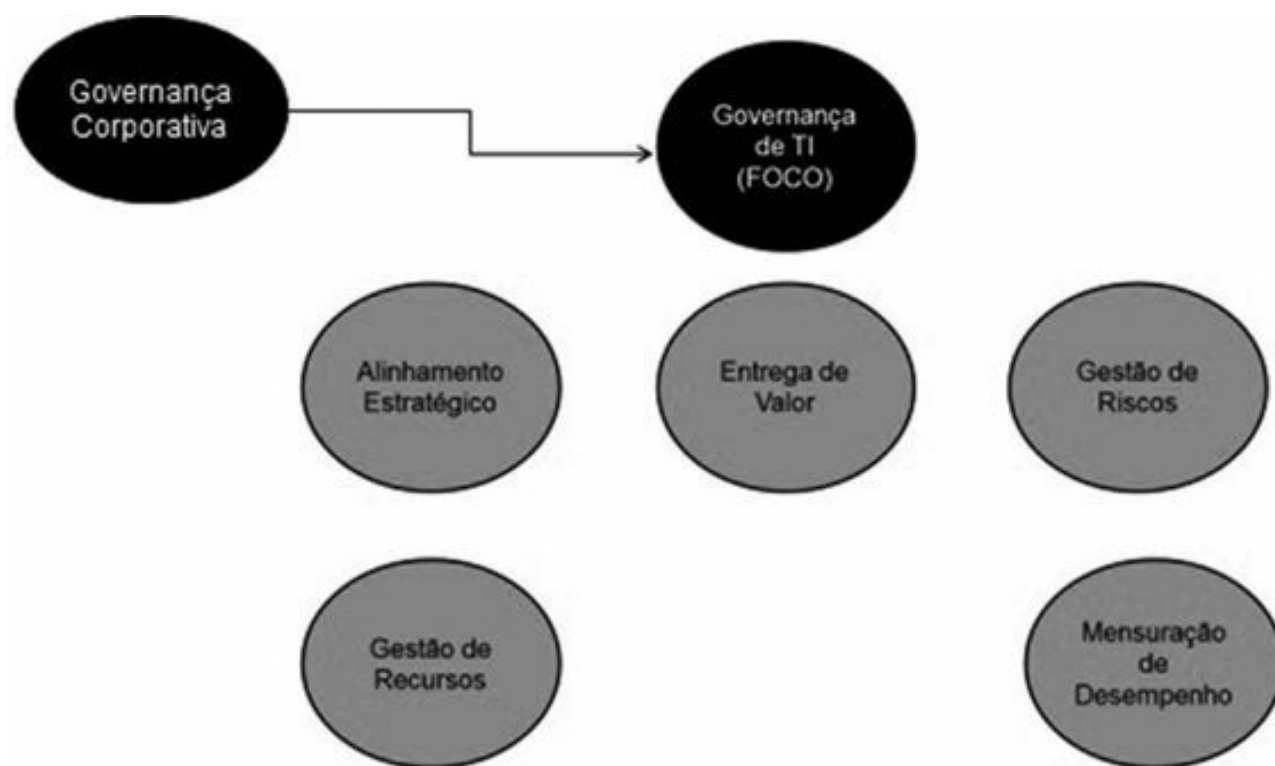


Figura 4-3 – Áreas Foco do COBIT

A escolha dos processos, em uma implantação de governança de TI baseada em COBIT, pode ser baseada na área-foco prioritária. Por exemplo, se fazer o alinhamento estratégico for a prioridade, pode-se escolher processos do COBIT que permitem realizar este alinhamento, deixando as outras áreas-foco para um segundo momento.

Uma parte também importante da Governança de TI é definir a estrutura de decisão. Weill e Ross (2006) criaram a matriz de arranjos de governança de TI que permite sistematizar as decisões de TI, considerando quais as principais decisões a serem tomadas (títulos das colunas) e quem as tomam (linhas).

As cinco principais decisões de TI, títulos das colunas da matriz, são:

- **Princípios de TI:** esclarece o papel de negócio da TI. Trata das declarações de alto nível sobre como a TI é e deve ser utilizada no negócio;
- **Arquitetura Empresarial:** define os requisitos de integração e padronização dos processos e sustenta o modelo operacional da organização. Trata da organização lógica de dados, aplicações e infraestrutura, definidas a partir de um conjunto de políticas, relacionamentos e opções técnicas adotadas para obter a padronização e a integração técnicas e de negócio desejadas;
- **Infraestrutura de TI:** determina os serviços de entrega e de suporte da TI. Trata dos

serviços de TI, coordenados de maneira centralizada e compartilhada, que fornecem a base para a capacidade de TI da empresa;

- **Necessidade de Aplicações de Negócio:** especifica as necessidades de aplicações, quer sejam adquiridas em formas de pacote ou desenvolvidas internamente;
- **Investimentos e Priorização de TI:** trata da escolha de que iniciativas financiar e quanto gastar. Trata de decisões sobre quanto e onde investir em TI, incluindo a aprovação de projetos e as técnicas de justificativa.

Na matriz de arranjos de governança de TI, os títulos das linhas listam um conjunto de arquétipos que identificam as pessoas envolvidas nas decisões de TI:

- **Monarquia de Negócio:** os altos gerentes.
- **Monarquia de TI:** os especialistas em TI.
- **Feudalismo:** cada unidade de negócio toma decisões independentes.
- **Federalismo:** combinação entre o centro corporativo e as unidades de negócio, com ou sem o envolvimento do pessoal de TI.
- **Duopólio:** o grupo de TI e algum outro grupo.
- **Anarquia:** tomada de decisões individual ou por pequenos grupos de modo isolado.

Esses arquétipos descrevem os arranjos decisórios que os autores citados encontraram em uma pesquisa com as maiores empresas americanas. O desafio sugerido pelo modelo é determinar quem deve ter a responsabilidade por tomar e contribuir com cada tipo de decisão de governança de TI. A **Figura 4-4** ilustra a matriz citada. As decisões de utilizar a nuvem devem obrigatoriamente obedecer a matriz de governança de TI.

As organizações possuem diferentes arranjos decisórios que muitas vezes não são explícitos. Estes arranjos estão intimamente ligados aos mecanismos de controle da informação e a sua cultura e, em muitas situações, para o bem da organização, precisam ser mudados. A mudança só é realmente efetivada com o apoio e envolvimento das lideranças do primeiro time de executivos que estão fora da organização de TI.

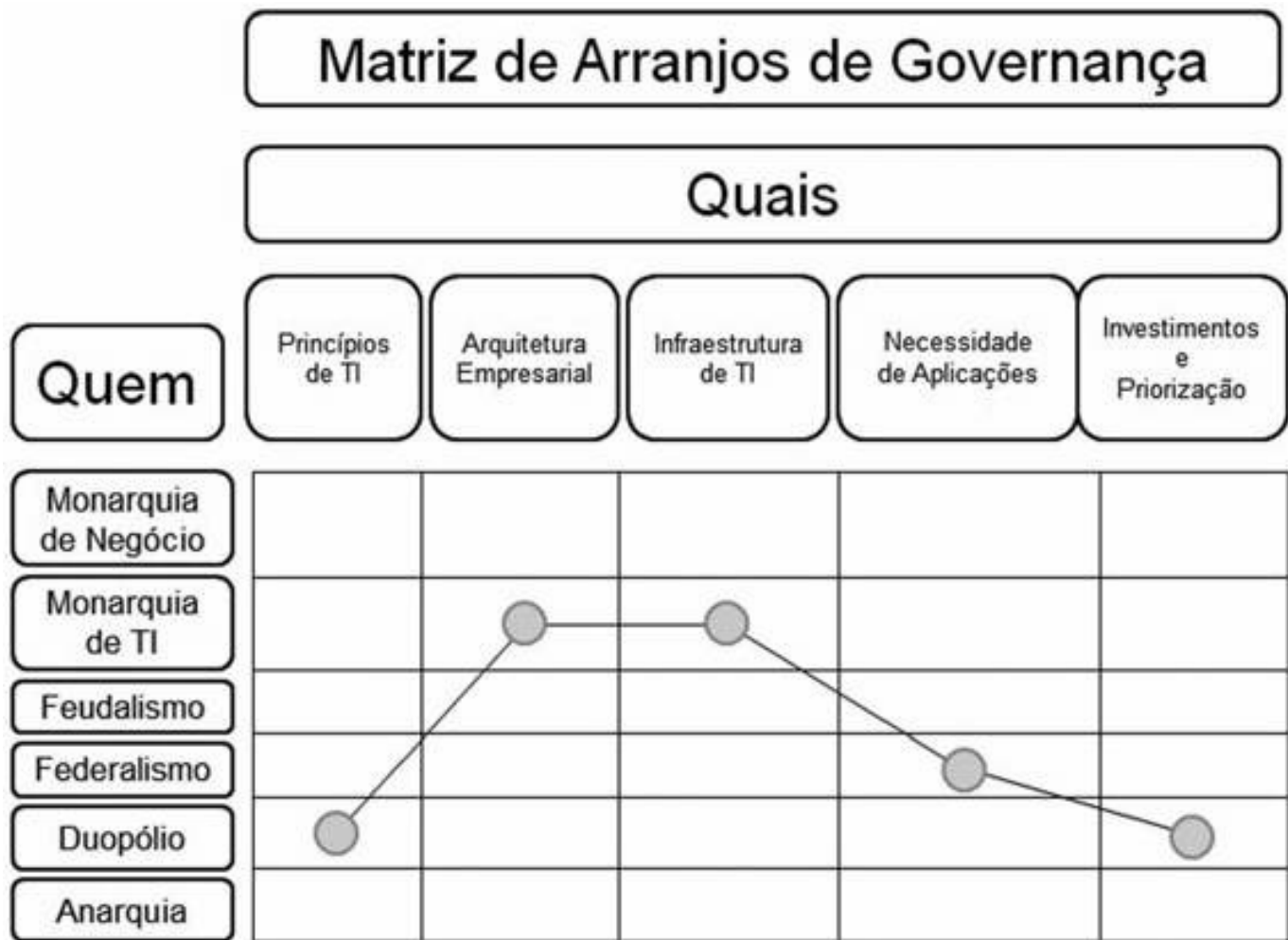


Figura 4-4 – Matriz de Arranjos de Governança de TI

Weill e Ross (2006) observaram que na maioria das empresas onde a governança de TI é considerada de boa qualidade as decisões referentes a princípios de TI e priorização de investimentos são baseadas em arranjos do tipo duopólio, onde as decisões pertencem ao grupo de TI e algum outro grupo. As decisões relativas a infraestrutura e arquitetura empresarial na maioria dos casos utilizavam um arranjo do tipo monarquia de TI.

4.3. Governança de TI e CLOUD COMPUTING

No caso de adoção da nuvem, pode-se assumir que esta decisão está ligada diretamente às decisões de arquitetura empresarial (relacionada diretamente com a arquitetura de TI) e infraestrutura de TI. Decisões compartilhadas podem ser uma saída para decisões de adoção da nuvem. Também a arquitetura empresarial influencia e é influenciada pela infraestrutura de TI. É mais comum que a arquitetura empresarial (AE), que sustenta o modelo operacional e, logicamente, a estratégia, influencie a infraestrutura de TI.

A decisão compartilhada para arquitetura empresarial e infraestrutura de TI permite tirar o peso da decisão sobre o diretor de TI, ao mesmo tempo em que força a participação dos outros executivos na decisão, fazendo com que eles saibam como a empresa funcionará neste novo modelo. Como a mudança tem diversas implicações, melhor que pessoas de negócio e de TI assumam suas

responsabilidades neste novo contexto.

A **Figura 4-5** ilustra esta sugestão.

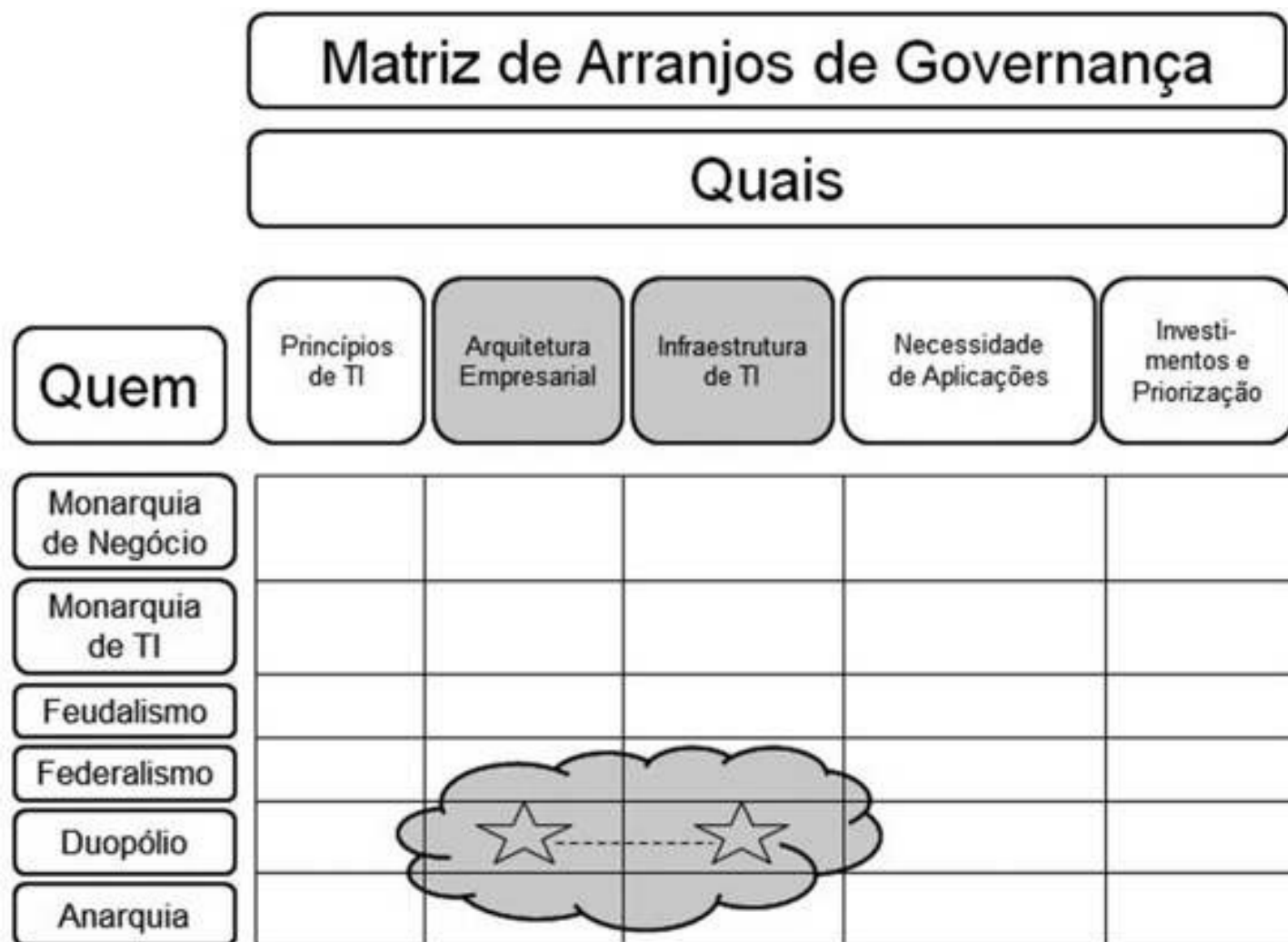


Figura 4-5 – Matriz sugerida para decisão de adoção de CLOUD COMPUTING

O que acontece com as empresas que possuem uma governança de TI efetiva? Segundo Weill e Ross (2006), têm lucros até 20% maiores do que empresas que buscam estratégias similares.

Como a nuvem afeta o departamento de TI

Depois de ter decidido optar pela nuvem, preparar-se para a transição para a nuvem será a próxima etapa. Um ponto chave é avaliar como a implantação afetará os ativos de TI, incluindo o pessoal de TI existente, e agir para garantir que a transição possa ser administrada sem contratempos.

Executar um diagnóstico faz parte da rotina de qualquer projeto de implantação de infraestrutura de TI bem-sucedido. Alguns fatores merecem consideração especial:

- **Padrões de segurança dos dados.** Transferir os dados críticos do negócio para a nuvem introduz o risco da perda de dados ou da exposição inadvertida de informações sigilosas. É necessário avaliar as necessidades de segurança dos dados e confirmar se o provedor adota medidas no local para atender aos padrões definidos.
- **Garantia do SLA.** O contrato de gerenciamento a ser firmado entre a organização e o

provedor de nuvem assume o formato baseado em SLA, que garante o nível de desempenho, disponibilidade e segurança que o provedor SaaS fornecerá e que regerá as ações do provedor – ou a indenização – caso este deixe de cumprir essas garantias. Verificar se os SLAs propostos estão em vigor e se as garantias prometidas são suficientes para atender suas necessidades.

- **Estratégia de migração.** Em um determinado momento, poderá ser desejado migrar de um aplicativo em SaaS, por exemplo, para outra solução, e por isso é importante estar em condições de tirar os dados existentes do aplicativo e transferi-los para outro. Perguntar ao futuro provedor de nuvem quais estratégias e procedimentos para migração de dados ele usa, inclusive se há provisões relativas à garantia para dados e códigos.
- **Exigências para integração interna.** É necessário confirmar se a migração para a nuvem atenderá todas as exigências funcionais e de integração de dados adotadas pela empresa.
- **Serviços de relatórios.** Considerando que a adoção da nuvem envolve abandonar o controle direto de alguns de seus dados, a emissão de relatórios exatos e úteis é especialmente importante. Saber quais são os serviços de relatório oferecidos pelo provedor e se são compatíveis com as exigências do negócio.
- **Impacto sobre as funções e as responsabilidades de TI.** Conforme mencionado anteriormente, adicionar serviços de nuvem ao mix de TI da empresa poderá provocar uma mudança fundamental no papel do departamento de TI, que vê alteradas as funções e responsabilidades com a implantação do novo modelo. A resistência a mudança será grande no próprio departamento de TI.
- **Impacto sobre o cumprimento dos regulamentos.** Pode ser importante solicitar um relatório do tipo SAS 70 do futuro provedor SaaS, mas também fazer uma análise detalhada para verificar se o provedor tem condições de cumprir suas próprias normas internas relativas às questões de privacidade, segurança de dados, etc. Por exemplo, se uma lei de privacidade local exigir que os dados pessoais financeiros dos clientes sejam sempre armazenados criptografados, o relatório SAS 70 do provedor divulgará se as práticas de armazenamento de dados desse provedor permitem utilizar tal tecnologia.

4.4. Terceirização da TI e CLOUD COMPUTING

4.4.1. Introdução

CLOUD COMPUTING privada, pública ou mesmo híbrida pode ser vista como uma forma de terceirização da TI. Mesmo no caso privado, a operação do DATACENTER interno pode estar nas mãos de terceiros, o que não é incomum, portanto, como um processo de terceirização (*outsourcing*). Lógico que a terceirização para um provedor externo com uma configuração de nuvem pública é muito mais complexa e demanda um esforço maior na definição da escolha do provedor e no

gerenciamento dos níveis de serviço. Toda a teoria e a prática sobre terceirização da TI já existente pode ser utilizada para ajudar na decisão de adoção de CLOUD COMPUTING.

4.4.2. Classificação da Terceirização de TI

A classificação da terceirização ajuda a entender de que forma acontece a terceirização na adoção de nuvem. A terceirização da TI pode acontecer de várias formas, destacando-se:

Quanto ao número de provedores utilizados:

- Terceirização com provedor único.
- Terceirização seletiva com conjunto de provedores.
- Terceirização com consórcio de provedores.
 - **Opção A:** provedores selecionados e geridos por provedor primário. Provedor primário selecionado e gerido pela organização contratante.
 - **Opção B:** provedores geridos e selecionados pelo contratante. Provedores selecionados pela organização contratante e supervisionados operacionalmente por um provedor integrador.

Quanto à estratégia de implantação:

A terceirização também pode ser vista pela ótica do ritmo de progressão da implantação onde se distinguem basicamente duas alternativas:

- **Terceirização total:** atividades são terceirizadas de uma única vez.
- **Terceirização incremental:** atividades são terceirizadas progressivamente.

Quanto ao tipo de relacionamento:

- **Parcerias estratégicas:** terceiro assume a responsabilidade por um conjunto integrado de operações do cliente.
- **Co-sourcing:** cliente e fornecedor compartilham a responsabilidade administrativa pelo sucesso do projeto.
- **Transação:** terceiro executa para o cliente um processo repetível e bem definido de TI ou de negócios por ele habilitados.

4.4.3. Benefícios da Terceirização da TI

No capítulo anterior foram mostrados os benefícios esperados com a adoção da CLOUD COMPUTING. Existem também benefícios mais gerais esperados com um projeto de terceirização (SAAD, 2006):

- Reduzir e controlar custos operacionais.
- Incrementar o grau de flexibilidade.

- Reduzir o prazo de disponibilização de novos produtos.
- Utilizar recursos especializados em áreas específicas.
- Melhorar a qualidade dos serviços de TI.
- Ganhar acesso às melhores práticas de indústria.
- Melhorar o retorno sobre bens, reduzir bens de capital, minimizar futuros investimentos de capital.
- Manter equipe atualizada tecnologicamente.
- Dar foco nas competências centrais.
- Compartilhar riscos.
- Obter injeção de recursos financeiros com a venda de bens.

Estes benefícios gerais podem ser obtidos com a adoção da nuvem, além dos específicos colocados no capítulo anterior.

4.4.4. Riscos da Terceirização de TI

No capítulo anterior mostraram-se os riscos esperados com a adoção da CLOUD COMPUTING. Existem também os riscos associados a um processo de terceirização de TI qualquer. SAAD (2006) sugere que estes riscos gerais em um processo de terceirização de TI podem ser avaliados nas fases de seleção do provedor, negociação do contrato e gestão do contrato.

- **Seleção do provedor de nuvem:** o processo de seleção de um provedor de nuvem não está isento de riscos. Os riscos deverão ser identificados, priorizados, controlados e tratados de forma a garantir que a escolha recaia sobre o provedor mais qualificado.
- **Negociação do contrato com o provedor de nuvem:** o processo de negociação do contrato com o provedor apresenta inúmeros elementos de risco, devendo o contrato incorporar controles e tratamento eficazes para os riscos identificados mais relevantes, evitando com isso o surgimento de crises ao longo da vida do contrato.
- **Gestão do contrato com o provedor de nuvem:** ao longo do ciclo de vida do contrato, tanto o provedor de nuvem como o cliente devem demonstrar competência para reduzir a exposição aos riscos identificados.

4.5. Governança de TI e Arquitetura Empresarial

A governança de TI pode ser exercida de duas formas. Quando o foco é o planejamento estratégico de TI ela assume principalmente o papel de alinhar os planos estratégicos da organização e de TI. Esta não é uma tarefa fácil, mas boa parte das organizações adota esta opção.

A outra forma é vincular a governança de TI aos projetos, encarados um de cada vez. Sabe-se

que sem a adoção de mecanismos de vinculação, os líderes de projeto agem isolados. Eles escolhem soluções que atendem as metas de projeto, mas as metas gerais de integração e padronização da empresa são ignoradas, e a arquitetura empresarial não passa de uma miragem. Governança de TI eficiente e gestão de projetos disciplinados podem ainda sim representar um envolvimento de TI ineficaz. (ROSS, WEILL e ROBERTSON, 2006).

A ideia é criar mecanismos de vinculação que cuidem para que os projetos construam o alicerce de execução da empresa (modelo operacional, arquitetura empresarial). Os mecanismos de vinculação devem permitir a construção do alicerce de execução com um projeto por vez.

Os autores citados sugerem três principais mecanismos de vinculação para o modelo de governança de TI:

- **Vinculações arquitetônicas:** estabelecem e atualizam padrões, examinam projetos e sua relevância e aprovam exceções.
- **Vinculações de negócios:** asseguram que metas comerciais se traduzam em metas de projeto.
- **Vinculação de alinhamento:** asseguram a comunicação e a negociação constantes entre questões de negócio e de TI.

Implementação da TI à moda antiga, ou seja, articulando a estratégia de negócios e alinhando a TI normalmente, torna a TI um gargalo. Novos projetos devem gerar valor agregado em noventa dias e não em quatro anos, horizonte trabalhado no planejamento estratégico. Em quatro anos tudo estará diferente.

O modelo baseado na arquitetura empresarial permite que a empresa se prepare para futuras iniciativas estratégicas sem saber quais são elas. A ideia do modelo é concentrar investimentos no fomento à integração e à padronização dos processos exigidas pelo modelo operacional.

4.6. Terceirização da TI e Arquitetura Empresarial

A ideia é utilizar a arquitetura empresarial para orientar a terceirização. As empresas terceirizam parte ou toda a TI por uma variedade de razões, incluindo redução de custos, mitigação de riscos, obtenção de capacidade variável, oportunidade de se concentrar nas competências centrais. Redução de custos e obtenção de capacidade variável são frequentemente apontadas como as maiores razões para a terceirização.

Uma forma de entender a relação entre a terceirização e a arquitetura empresarial é relacionar os modelos de terceirização, sob a ótica do relacionamento, com os estágios da arquitetura empresarial. A **Tabela 4-1** ilustra esta relação.

Tabela 4-1 – Terceirização versus Estágios de Arquitetura Empresarial

	Silos de Negócio	Tecnologia Padronizada	Núcleo Otimizado	Modularidade dos Negócios
O que terceirizar	Processos facilmente isolados	Administração da infra de TI	Gestão de Projetos de Grandes Sistemas	Design e Operação de Processos com tecnologia de apoio
Relacionamento ideal	Terceirização transacional com enfoque restrito	Parceria Estratégica	Aliança de Co-sourcing	Terceirização Transacional
Objetivos	Redução de Custos	Disciplina de gestão da TI, redução de custos, redução de riscos, enfoque administrativo	Transferência de Tecnologia Disciplina e reengenharia de processos, enfoque administrativo, custo eficiência, capacidade variável, partilha de riscos	Agilidade estratégica, aproveitamento da pericia em processos e TI para processos de classe mundial Capacidade variável Enfoque administrativo Custo eficiência Partilha de riscos

Fonte: Enterprise Architecture as Strategy. Harvard Business School Press, 2006.

Existe uma relação entre terceirização e estágio arquitetônico. Diferentes relações de terceirização são adequadas a diferentes estágios. Ross, Weill e Robertson (2006) sugerem que a arquitetura empresarial não deve ser terceirizada. O que deve ser terceirizado é a parte da TI que suporta a construção do modelo operacional.

4.7. Conceito na Prática: CLOUD COMPUTING no Governo Americano

O CIO do governo americano recentemente lançou um documento na web intitulado *Federal Cloud Computing Strategy*, que estimula todo o governo federal americano a utilizar soluções de nuvem. Segundo este documento, a TI no governo federal sofre atualmente dos seguintes males:

- Baixo uso.
- Demanda fragmentada por recursos.
- Sistemas duplicados.
- Ambientes pouco gerenciáveis.

- *Leadtime* elevado para aquisições.

A ideia central do documento é aumentar a eficiência dos sistemas de TI através da adoção de soluções de nuvem e assim melhorar o impacto nos cidadãos. O orçamento de TI que poderia ser migrado para serviços de nuvem, segundo o documento, seria de US\$ 20 bilhões (US\$ 80 bilhões foi o orçamento total anual de 2010).

A utilização da nuvem no governo americano deve obedecer ao documento inicial lançado, intitulado Cloud First Policy (www.whitehouse.gov). O documento tem intenção de acelerar o uso de soluções de CLOUD COMPUTING pelas agências americanas antes destas realizarem novos investimentos.

Os modelos de implementação utilizados pelo governo seriam os modelos privado, público e híbrido.

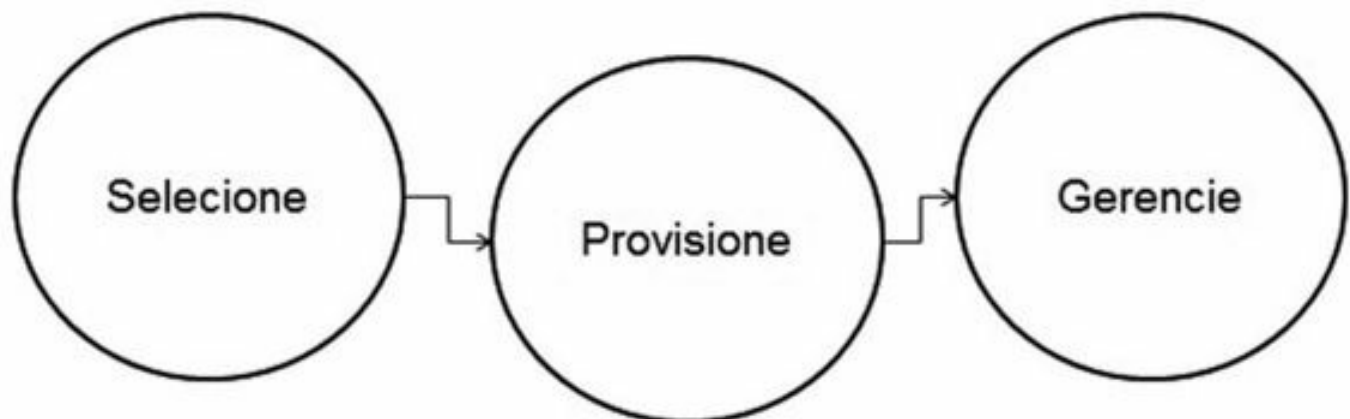
As estratégias do governo podem ser resumidas em:

- Articular benefícios, considerações e *trade-offs*.
- Propiciar um *framework* de decisão.
- Oferecer dicas sobre os recursos de implementação.
- Identificar atividades e responsabilidades para catalisar a adaptação.

Os benefícios a serem alcançados com as soluções de CLOUD COMPUTING foram divididos em três grupos:

- **Eficiência:** trata da melhoria do uso dos ativos, consolidação, melhorar produtividade no desenvolvimento de aplicações.
- **Agilidade:** melhoria do uso da capacidade, melhoria do tempo de resposta a solicitações urgentes, compra baseada em serviço.
- **Inovação:** melhoria da cultura, melhor *link* com tecnologias emergentes, mudança de foco de posse de ativos para gerência de serviços.

A **Figura 4-6** ilustra o framework sugerido pelo governo americano para apoiar a decisão de adoção da arquitetura CLOUD COMPUTING.



As fases são detalhadas a seguir:

- **Seleção:** identificar que serviços (valor versus prontidão) podem ir para a CLOUD e quando.
- **Provisione:** agregar demanda no nível de departamento quando possível, estabelecer interoperabilidade e integração com o portfólio de TI, contratos devem representar as demandas das agências, realizar valor.
- **Gerencie:** mudança de ativos para serviços. Construir novas habilidades, monitorar SLAs, reavaliar fornecedores e modelos de serviço procurando reduzir riscos e otimizar benefícios.

Recentemente a Amazon AWS lançou a região AWS GovCloud (US) para atender a esta demanda do governo americano. Vale ressaltar que o DATACENTER da nuvem AWS GovCloud é exclusivo para uso por empresas governamentais americanas. Esta região (DATACENTER) é específica para acesso de cidadãos americanos e atende a requisitos de compliance do governo americano. A [Figura 4-7](#) ilustra as regiões contempladas pelo modelo Amazon AWS.

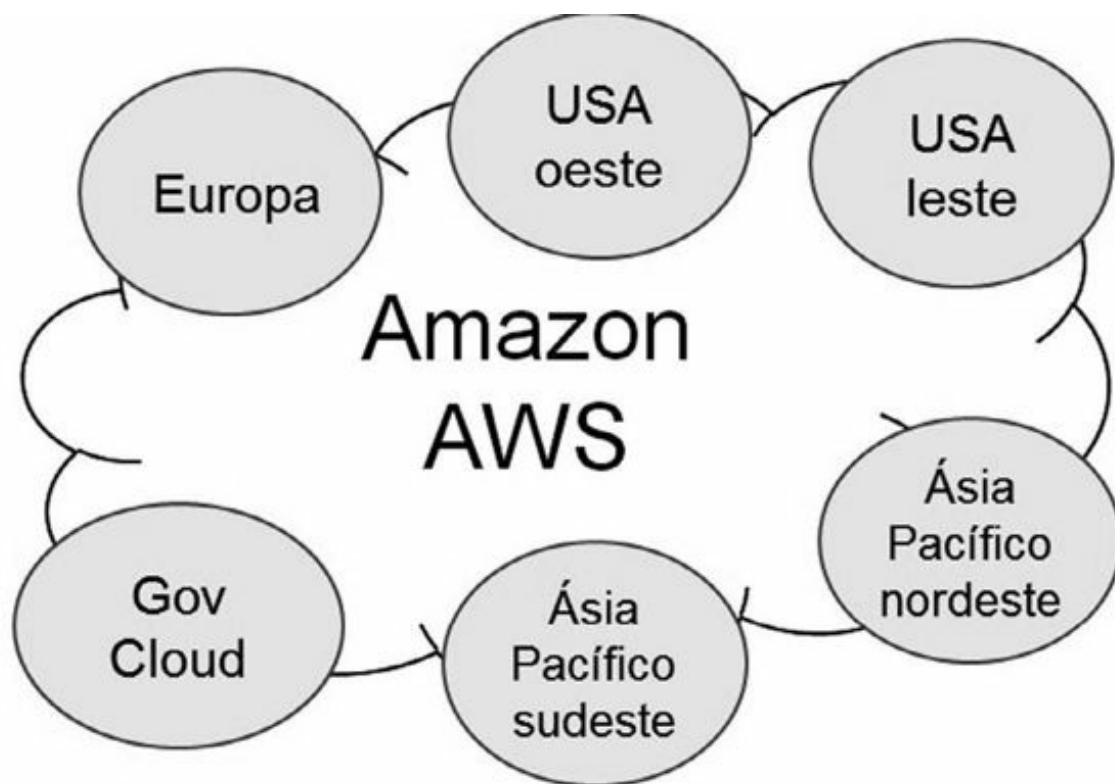


Figura 4-7 – Seleção de Serviços para a Migração para a nuvem
(Fonte: Federal CLOUD COMPUTING Strategy, Governo Americano, 2011)

4.8. Seleção do Provedor de CLOUD COMPUTING

Selecionar um provedor de CLOUD COMPUTING requer uma série de passos importantes.

Inicialmente deve-se decidir se vale a pena adotar serviços de nuvem. Depois são definidos os modelos de serviços a serem utilizados e as formas de implementação. O próximo passo é escolher o fornecedor (provedor) da solução. As opções de CLOUD COMPUTING são resumidas na [Figura 4-8](#).

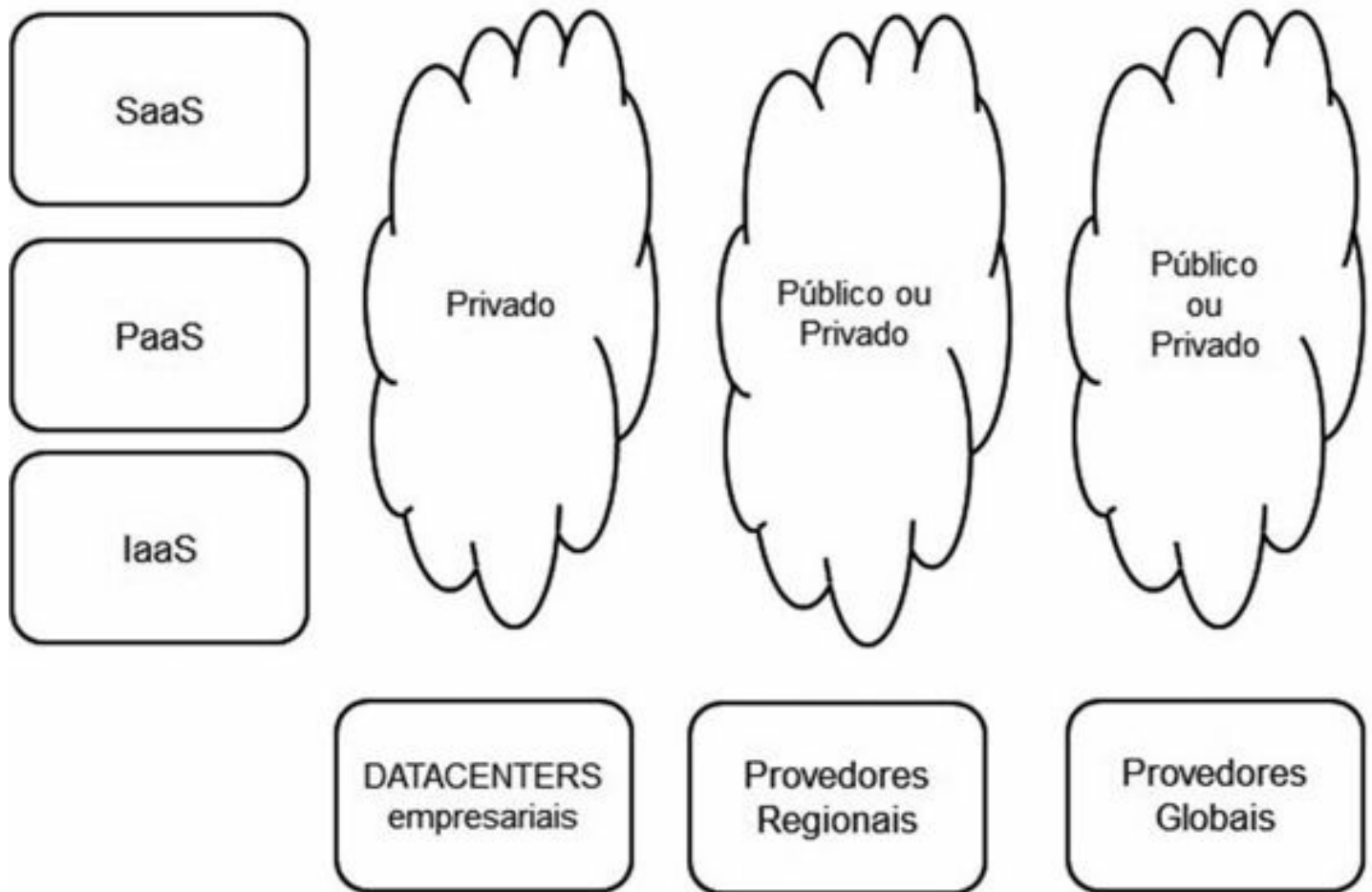


Figura 4-8 – Decisões de CLOUD COMPUTING

Cada provedor possui uma oferta que precisará ser comparada. Dependendo do tipo de provedor (Global, Regional ou mesmo um provedor interno para DATACENTERS empresariais), a oferta poderá variar bastante. Em certas situações será necessário adquirir serviços de mais de um provedor.

Saad (2006) sugere uma sequência de passos para a seleção de provedores de terceirização que pode muito bem ser adaptada aos provedores de CLOUD COMPUTING. Os passos para a escolha do(s) provedor(es), sugeridos por ele, são descritos a seguir:

- Pesquisa de mercado.
- Visitas a clientes deste tipo de terceirização.
- Pré-qualificação dos provedores.
- Envio de solicitação de proposta de serviços (RFP).
- Avaliação das propostas deve considerar o peso dos critérios e diferentes graus.

- Verificação das competências alegadas.
 - ☐ Expertise.
 - ☐ Metodologias.
 - ☐ Tecnologias.
 - ☐ Ferramentas.
 - ☐ Processos.
 - ☐ Posicionamento.
 - ☐ Inovação.
 - ☐ Experiência na Indústria.
 - ☐ Desempenho.
 - ☐ Suporte.
 - ☐ Serviços.
 - ☐ Treinamento.
- Verificação da capacidade.
 - ☐ Solidez financeira.
 - ☐ Reputação.
 - ☐ Recursos de infraestrutura.
 - ☐ Sistema gerencial.
 - ☐ Amplitude do portfólio de serviços.
- Verificação da dinâmica de relacionamento.
 - ☐ Adequação cultural.
 - ☐ Estratégia.
 - ☐ Flexibilidade.
 - ☐ Gestão do relacionamento.
 - ☐ Relação do porte cliente/provedor.
 - ☐ Importância relativa do cliente.
- Avaliar eficácia e competitividade da solução.
 - ☐ Adequação aos requisitos.
 - ☐ Grau de inovação.
 - ☐ Grau de risco.
 - ☐ Gestão de riscos.
 - ☐ Compartilhamento de riscos.

- Garantias.
- Proposta Financeira.
- Investimento pelo contratante.
- Flexibilidade para alteração do escopo.
- Prazo de implantação.
- Duração mínima do contrato.
- Termos e condições.
- Requisitos de RH.

Com base nos atributos e subatributos citados, se daria a escolha do provedor. Logicamente, é necessário pontuar estes atributos e subatributos, considerando a cultura e a realidade da empresa cliente.

4.9. Referências Bibliográficas

- CLOUD COMPUTING: Business Benefits with Security, Governance and Assurance Perspectives. ISACA, an ISACA Emerging Technology White Paper, 2009.
- Frederick Chong & Gianpaolo Carraro. Software como Serviço (SaaS): Uma perspectiva corporativa, Microsoft MSDN, 2007.
- Greenberg, Albert; Hamilton, James; Maltz, David A.; Patel, Parveen. The Cost of Cloud: Research problems in DATACENTER Networks, 2009.
- O papel da segurança na computação em nuvem confiável. RSA, 2009.
- Ross, Jeanne; Weill, Peter; Robertson, David C. Enterprise Architecture as Strategy. Harvard Business School Press, 2006.
- Saad, Alfredo C. Terceirização de Serviços de TI. Brasport, 2006.
- Saten, John. Deliver CLOUD Benefits Inside your Walls, Forrester, 2009.
- The Benefits and Risks of CLOUD Platforms: A Guide for Business Leaders, Microsoft, 2011.
- The Economics of the CLOUD. Microsoft, November, 2010.
- The NIST Definition of CLOUD COMPUTING (Draft), NIST, January 2011.
- Varia, Jinesh. Architecting for the Cloud: Best Practices, Amazon Web Services, January, 2010.
- Veras, Manoel. DATACENTER: Componente Central da Infraestrutura de TI, Brasport, 2009.
- Veras, Manoel. VIRTUALIZAÇÃO: Componente Central do DATACENTER, Brasport, 2011.
- Weill, Peter e Ross, Jeanne W. Conhecimento em TI. Tradução, M. Books, 2010.

PARTE II: INFRAESTRUTURA DE NUVEM

5. DATACENTER: Aspectos Gerais

5.1. Introdução

Um DATACENTER (ou DATA CENTER) é um conjunto integrado de componentes de alta tecnologia que permitem fornecer serviços de infraestrutura de TI de valor agregado, tipicamente processamento e armazenamento de dados, em larga escala, para qualquer tipo de organização. Os DATACENTERS e suas conexões formam a infraestrutura da nuvem, quer seja pública ou privada.

No caso da nuvem privada hospedada internamente, a empresa precisa construir o seu DATACENTER com os recursos que o façam funcionar em nuvem, incluindo a VIRTUALIZAÇÃO. Existem limitações neste modelo.

No caso da nuvem pública Microsoft, por exemplo, existem hoje seis DATACENTERS em operação, conforme ilustra a **Figura 5-1**. A localização em três grandes continentes não é por acaso. Questões de latência, segurança e *compliance* definem a localização dos dados e processamento das aplicações das organizações que estão na nuvem Microsoft:

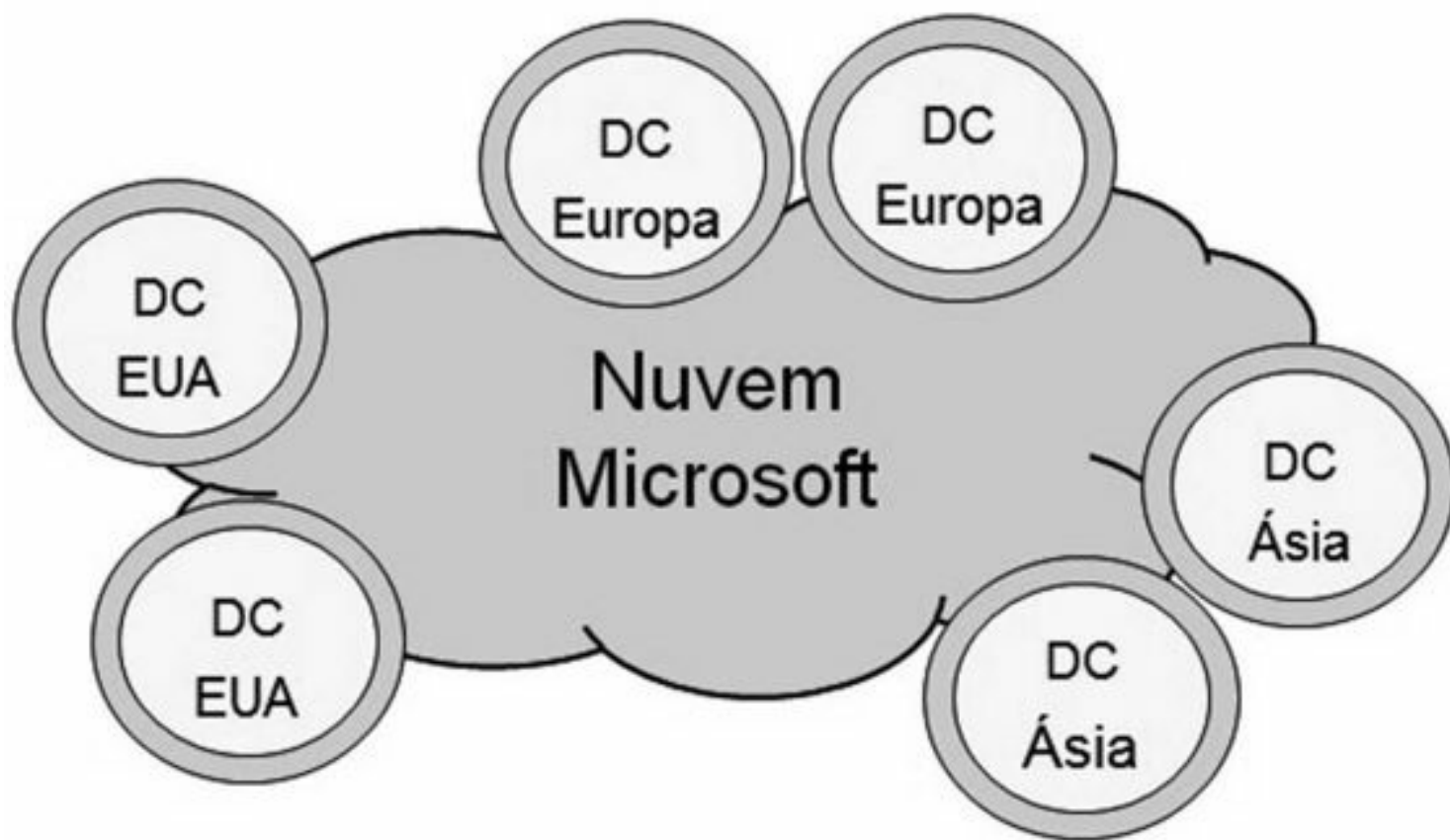


Figura 5-1 – CLOUD e DATACENTERS Microsoft

O componente central da CLOUD COMPUTING de qualquer organização é o DATACENTER, e toda organização de alguma forma possui um DATACENTER ou utiliza serviços de um DATACENTER. Os DATACENTERS são alimentados por energia e conectados ao mundo externo por uma rede de telecomunicações e servem a carga de TI. A **Figura 5-2** ilustra um típico DATACENTER.

Em um projeto de DATACENTER a carga de TI define as instalações e a energia a ser adquirida da concessionária. A utilização da TI define a banda necessária do sistema de telecomunicações.

Os DATACENTERS podem ser divididos em dois grandes grupos: os DATACENTERS empresariais (eDC) construídos normalmente dentro das instalações da empresa que o utiliza e os DATACENTERS de INTERNET (iDC), responsáveis pelos serviços de nuvem para terceiros.



Figura 5-2 – DATACENTER

A grande confusão que muita gente faz é achar que só existe DATACENTER nos provedores de serviço de Internet. Nos Estados Unidos (EUA), por exemplo, boa parte das organizações de médio e grande porte possui um DATACENTER próprio. O IDC apontou que nos EUA em 2008 já existiam cerca de **2,5 milhões** de DATACENTERS.

O mercado de DATACENTERS deverá crescer nos próximos anos a taxas muito expressivas. Estudo feito por Christian Belady da Microsoft, em 2011, mostra que em 2020 o mercado de DATACENTERS deverá ser de US\$ 218 bilhões, mesmo considerando a redução do preço do DATACENTER por WATT ocorrido pelo uso de novas tecnologias e da adoção da modularidade em novos projetos.

Para efeito didático, os DATACENTERS podem ser divididos em três grandes blocos: instalações (inclui instalações físicas, equipamentos de energia e refrigeração), gerenciamento e a TI, conforme ilustra a **Figura 5-3**.

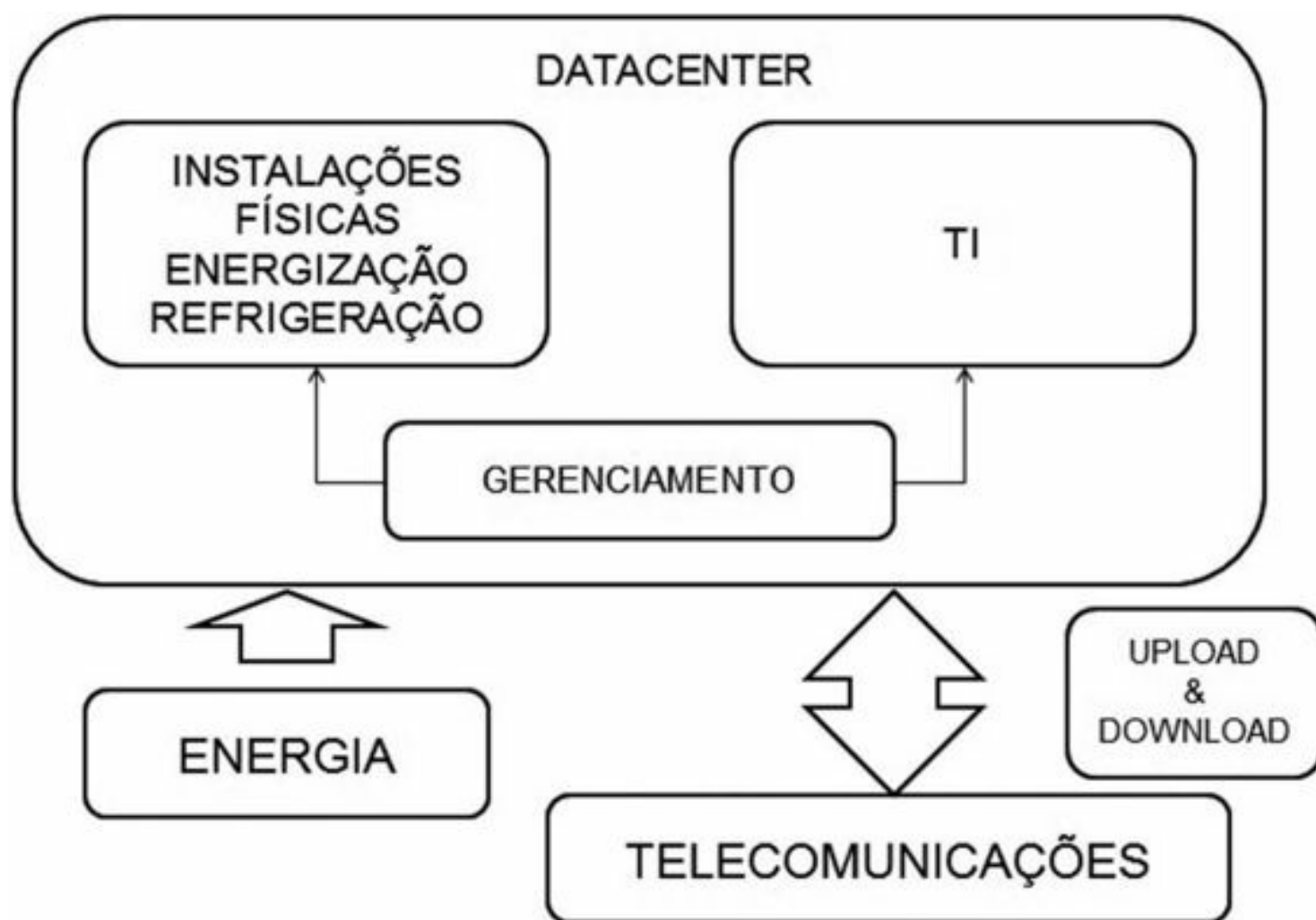


Figura 5-3 – Componentes do DATACENTER

As organizações estão repensando a arquitetura e o planejamento do DATACENTER, devido ao aumento das demandas de processamento e armazenamento e também devido ao aumento das normas e da fiscalização rigorosa sobre a recuperação destas estruturas em caso de desastres. Estas novas demandas acabam provocando grande impacto nos níveis de serviço fornecidos e nos custos das operações dos DATACENTER.

A consolidação de muitos DATACENTERS em poucos é uma resposta a estas novas demandas e só se torna possível devido aos grandes avanços dos últimos dez anos em disponibilidade e na confiabilidade de banda da rede de telecomunicações e da tecnologia de Internet.

A consolidação do DATACENTER envolve a migração de todos os ativos de TI, pessoas e os aplicativos que recebem suporte, portanto, um planejamento abrangente é crítico para que o negócio continue a funcionar sem interrupção.

Novos DATACENTERS estão sendo construídos considerando o ganho de escala e a busca da eficiência energética como determinantes do projeto. A localização é determinada principalmente pelas condições de fornecimento de energia, telecomunicações e o clima. O tamanho destas estruturas pode ser visto na [Tabela 5-1](#).

Tabela 5-1 – Novos DATACENTERS

Companhia	Mês/Ano	Localização	Tamanho (sq. feet)	Custo (M US\$)
Facebook	Fev/2010	Oregon	307.000	N/A
Microsoft	Set/2009	Dublin	N/A	500
National Security Admin	Jul/2009	Utah	1.000.000	2.000
I/O	Jun/2009	Phoenix	538.000	N/A
Apple	Mai/2009	Maiden	500.000	1.000

Fonte: The Economics of the Cloud, Microsoft, 2010 (adaptado por Manoel Veras)

O papel principal dos componentes do DATACENTER no novo cenário é possibilitar o atingimento do nível de serviço adequado para cada aplicativo. A ideia central de um projeto de DATACENTER é oferecer níveis de serviço de acordo com a criticidade do aplicativo ao mesmo tempo em que é eficiente em termos energéticos.

Os níveis de serviço de DATACENTER dependem dos componentes do DATACENTER. A meta central de um projeto de DATACENTER é obter resiliência, que em TI significa atender a demanda de negócios de maneira efetiva, reduzir o custo total de propriedade (TCO) e tornar o negócio flexível. Os critérios a serem considerados em um projeto de DATACENTER para o atingimento desta meta são:

- Desempenho.

- Disponibilidade.
- Escalabilidade.
- Segurança.
- Gerenciabilidade.

Estes critérios devem ser utilizados no dimensionamento dos diversos dispositivos que compõem o DATACENTER e definem a qualidade dos serviços a serem providos para as aplicações. Estes serviços também devem ser projetados para funcionar em conjunto, ou seja, integrados, e atender as demandas das aplicações que, por sua vez, atendem as demandas dos processos de negócio.

Uma visão mais adequada do DATACENTER deve pensá-lo em termos de serviços. A qualidade dos dispositivos tecnológicos utilizados irá influenciar os níveis de serviço que serão entregues às aplicações e aos processos de negócio.

A **Figura 5-4** ilustra de maneira didática os principais serviços providos pelo DATACENTER, separando os serviços de TI dos componentes externos do DATACENTER como instalações, energia, refrigeração e gerenciamento físico para facilitar o entendimento. Alguns serviços acabam dependendo de outros e alguns serviços acabam completando outros serviços.

A VIRTUALIZAÇÃO é o principal serviço de TI do DATACENTER e influenciará o dimensionamento dos demais serviços oferecidos. O correto projeto da estrutura de TI virtualizada é a chave do bom projeto de DATACENTER.

Justificar um projeto de VIRTUALIZAÇÃO para uma infraestrutura nova ou mesmo já existente é uma tarefa fácil. As vantagens são inúmeras e o custo total de propriedade de uma solução virtualizada é facilmente demonstrado com a utilização de ferramentas disponibilizadas pelos fabricantes dos softwares de VIRTUALIZAÇÃO. Importante ressaltar que a VIRTUALIZAÇÃO permite aperfeiçoar o uso dos recursos de infraestrutura, mas não tem nenhum efeito direto na arquitetura da aplicação.

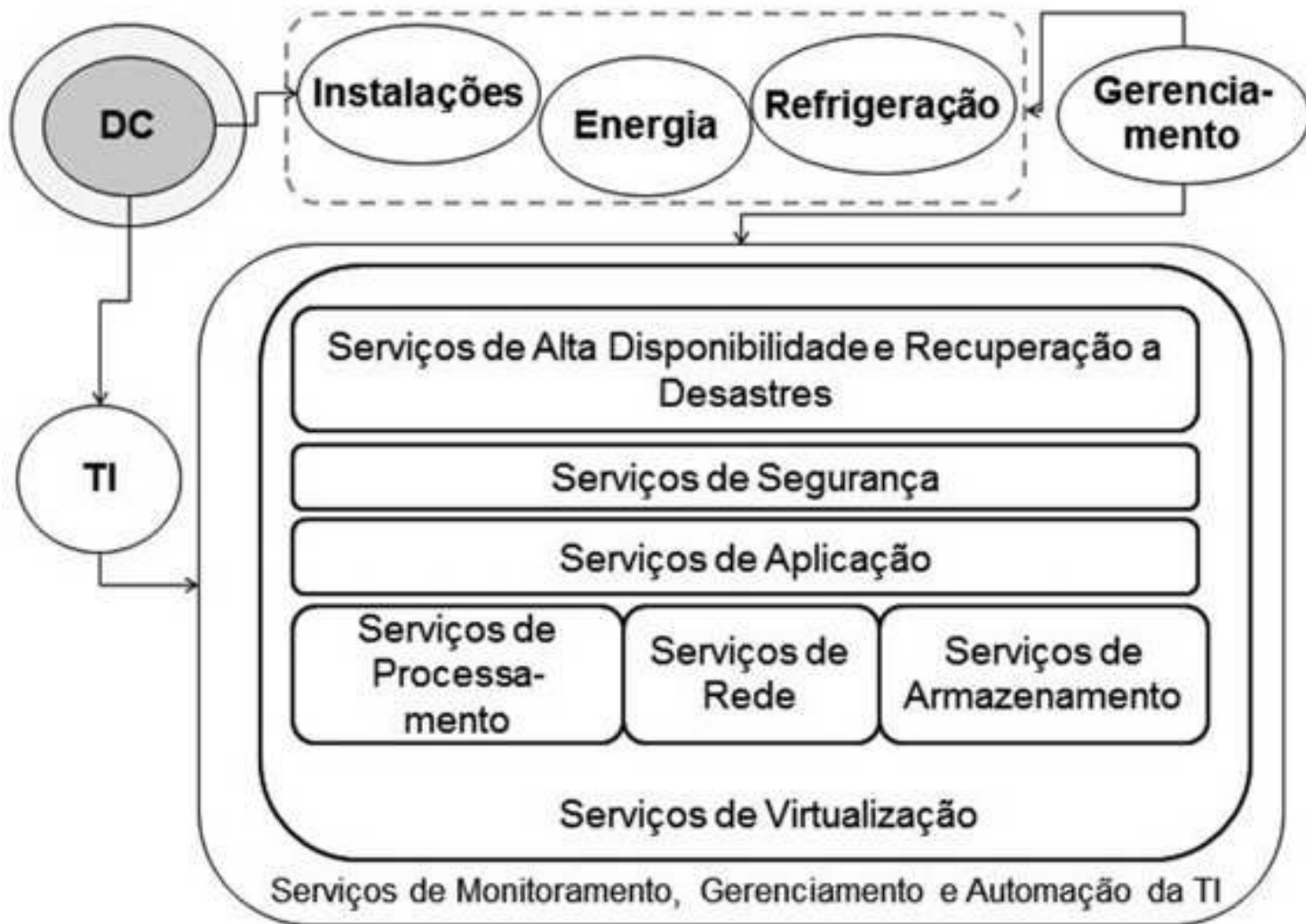


Figura 5-4 – Serviços de DATA CENTER

Os principais serviços do DATA CENTER que permitem atingir os níveis de serviço necessários para as aplicações são:

- **Serviços de rede:** envolvem a conexão entre os componentes internos e deles com o mundo exterior. As conexões são feitas mediante a arbitragem dos switches. Os switches são dispositivos importantes de um projeto de conectividade IP e são utilizados em várias camadas para definir a hierarquia do DATA CENTER.
- **Serviços de segurança:** serviços de segurança envolvem os serviços de *firewall*, que permitem o controle da navegação do usuário, negando acesso a aplicações e sites desnecessários, além de incorporar um sistema que permite aos gestores monitorar as atividades de cada usuário. Incluem também serviços providos por sistemas de detecção de intrusão e prevenção e filtragem de pacotes.
- **Serviços de processamento:** os serviços de processamento respondem diretamente pelo desempenho do DATA CENTER. Os dispositivos envolvidos são os servidores, sistemas operacionais e processadores. Os servidores podem utilizar diversas arquiteturas, mas a tendência maior é a de utilizar a arquitetura x86.
- **Serviços de armazenamento:** envolvem o armazenamento de dados em unidades de

storage. Redes de armazenamento e dispositivos de conexão ao *storage* são partes importantes da arquitetura. Também os níveis de serviço para disponibilidade e segurança são dependentes deste serviço.

- **Serviços de VIRTUALIZAÇÃO:** os serviços de VIRTUALIZAÇÃO permitem que servidores físicos rodem diversas aplicações em diferentes sistemas operacionais otimizando a utilização dos recursos de processamento e memória. Sistemas operacionais de VIRTUALIZAÇÃO e suas funcionalidades são aspectos importantes de serem tratados.
- **Serviços de aplicação:** os serviços de aplicação envolvem principalmente *load-balancing*, *secure socket layer (SSL)* *offloading* e *caching*.
- **Serviços de alta disponibilidade (*High Availability – HA*) e recuperação de desastres (*Disaster Recovery – DR*):** são serviços para obtenção da alta disponibilidade e recuperação de desastres, incluindo a extensão da SAN, seleção do site de contingência, interconectividade. Tratam também de políticas, softwares e dispositivos de backup e *restore* e replicação.
- **Serviços de monitoramento, gerenciamento e automação:** envolvem toda a malha de monitoramento, incluindo o NOC (*Networking Operation Center*), desde o hardware até os aspectos de automação de patches (correções) de sistema operacional. O gerenciamento deve possibilitar uma operação assistida 24x7 e executada até remotamente. O planejamento e o controle da produção devem garantir a execução dos processos de TI como paradas programadas, execução de *jobs* e rotinas periódicas.

A **Figura 5-5** ilustra a relação entre os serviços de DATACENTER e os serviços de infraestrutura de TI, normalmente referenciados por melhores práticas sugeridas pela biblioteca ITIL. O nível de serviço percebido pelo usuário do processo depende do nível de serviço da aplicação, que, por sua vez, depende do nível de serviço da infraestrutura, onde os serviços de DATACENTER estão incluídos. Assim, os componentes que você escolhe para montar a TI do DATACENTER têm relação direta com o nível de serviço que o usuário percebe.

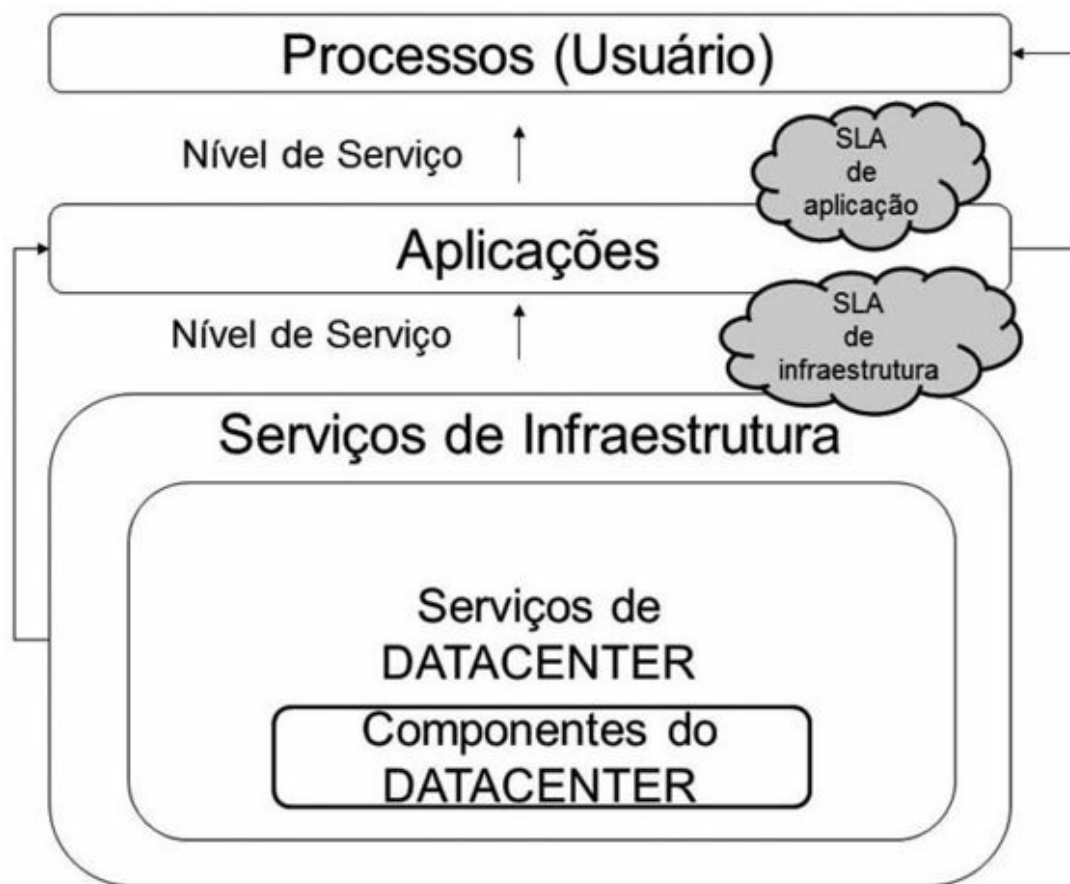


Figura 5-5 – Serviços de infraestrutura de TI e serviços de DATACENTER

Os serviços de DATACENTER são interdependentes e devem ser implementados de forma integrada, de modo a possibilitar o atingimento do nível de serviço, entregue à aplicação. Ou seja, o projeto de um DATACENTER deve considerar todos estes serviços de forma a obter os níveis de serviço necessários às aplicações e consequentemente ao negócio. Falar de nível de serviço por aplicação é uma tarefa árdua. Em geral o provedor de serviços tem dificuldades de propor um acordo de nível de serviço com base na aplicação, pois ele mesmo, na maioria das vezes, não consegue monitorar e portanto garantir o SLA por aplicação.

Os DATACENTERS, a rede de comunicação e as aplicações respondem pelo nível de serviço da nuvem.

5.2. UPTIME Institute

5.2.1. Introdução

O Uptime Institute (www.uptimeinstitute.org) é uma organização americana fundada em 1993 com foco em promover o aumento da confiabilidade e da disponibilidade de DATACENTERS. Uma das suas principais contribuições foi a criação da classificação de DATACENTERS em camadas (TIERS) que originou a norma TIA-942.

5.2.2. Norma TIA-942

A Norma TIA-942 (*Telecommunications Infrastructure for DATACENTERS*) de 2005 define os requisitos mínimos da infraestrutura de telecomunicações de DATACENTERS e salas de computadores. A norma vale para qualquer tamanho de DATACENTER e para DATACENTERS empresariais (*single tenant* ou um inquilino por aplicação) e DATACENTERS na Internet (*multitenant* ou multi-inquilino por aplicação).

A norma TIA-942 trata das instalações do DATACENTER e considera diversos aspectos do DATACENTER, incluindo o layout e espaço físico necessário, seleção do site, infraestrutura de cabeamento, aspectos mecânicos e elétricos.

A classificação em camadas por níveis de disponibilidade é a principal contribuição da norma.

O desempenho e a disponibilidade do DATACENTER podem ser descritos por uma série de termos como confiabilidade, disponibilidade e MTBF, normalmente difíceis de serem calculados. Uma alternativa a estes termos é categorizar o DATACENTER em termos de camadas (TIERS) de níveis críticos. Existe sempre um compromisso entre o custo total de propriedade (TCO) e a escolha de uma camada de criticidade. Ou seja, quanto mais sofisticado for o DATACENTER, menor o *downtime*, maior é o seu custo total de propriedade.

No projeto do DATACENTER, a etapa de planejamento requer muito cuidado, pois é onde normalmente se encontram os maiores erros. Deve-se construir uma especificação resultante das preferências e limitações do projeto e tendo como referência a norma TIA-942 e os aspectos de eficiência energética.

A APC no relatório técnico 122 – *Guidelines for Specifying DATACENTER Critically/ Tier Levels* aponta os métodos de classificação de DATACENTER mais conhecidos que são o *Uptime Institute's Tier Performance Standards*, TIA 942 e *Syska Hennessy Group's Criticality Levels*.

- **Uptime Performance Standard:** criado em 1995, é muito utilizado em projetos de DATACENTERS mas não especifica detalhes de projeto. A criticidade deve ser sempre obedecida em todos os critérios especificados. Define os níveis I a IV de disponibilidade.
- **ANSI/EIA/ TIA-942:** baseado nas quatro camadas do *Uptime Performance Standard*, já está na revisão 5, de 2005. Será detalhada um pouco mais adiante, por ser a mais utilizada. Define as camadas TIER 1 a TIER 4 de disponibilidade dos DATACENTERS.
- **Syska Hennessy Group:** baseado em dez camadas que se relacionam com as quatro camadas do *Uptime Performance Standard*. Os níveis de disponibilidade variam de 1 a 10.

A norma TIA-942 permite definir o nível de criticidade do DATACENTER e é sem dúvida o padrão utilizado para classificar DATACENTERS no mundo. A utilização da norma deve considerar se está se tratando de um novo projeto (chamado *greenfield* nos EUA) ou um DATACENTER já existente (chamado *retrofit* nos EUA). Para DATACENTERS de provedores é normal que o nível de criticidade seja elevado, pois os custos envolvidos para a obtenção de um alto nível de criticidade são diluídos. No caso de DATACENTERS empresariais esta é uma questão complexa, pois normalmente os custos para obtenção de um alto nível de criticidade são muito altos.

A norma TIA 942 indica os requisitos desde a construção até a pronta ativação do DATACENTER. Esta norma é baseada em um conjunto de outras normas relacionadas a seguir:

- **ASHRAE:** trata da refrigeração.
- **TIA/EIA 568:** estabelece um sistema de cabeamento para aplicações genéricas de telecomunicações em edifícios comerciais. Permite o planejamento e a instalação de um sistema de cabeamento estruturado.
- **TIA/EIA 569:** norma que trata dos encaminhamentos e espaços.
- **TIA/EIA 606:** providencia um esquema de administração uniforme independente das aplicações.
- **TIA/EIA 607:** trata das especificações de aterramento e links dos sistemas de cabeamento estruturado em prédios comerciais. Providencia especificações sobre aterramento e links relacionados à infraestrutura de telecomunicações do edifício.

A classificação em camadas TIA-942 é muito utilizada para comparação de DATACENTERS, estimativa de custo e desempenho.

IMPORTANTE: O *rating* da camada só é mantido se os aspectos chaves da camada são considerados. Ou seja, não adianta ter um DATACENTER robusto em termos elétricos e um sistema inadequado de refrigeração, pois, para efeito de classificação, o que vale é o sistema inadequado de refrigeração.

Os principais componentes da arquitetura do DATACENTER segundo a norma TIA-942 são descritos abaixo:

- ***Entrance Room (ER):*** sala de entrada. Espaço de interconexão entre o cabeamento estruturado do DATACENTER e o cabeamento vindo das operadoras de telecomunicações.
- ***Main Distribution Area (MDA):*** área onde se encontra a conexão central do DATACENTER e de onde se distribui o cabeamento estruturado. Inclui roteadores e o *backbone*.
- ***Horizontal Distribution Area (HDA):*** área utilizada para conexão com área de equipamentos. Inclui o *cross connect* horizontal e equipamentos intermediários. Incluem switches LAN/SAN/KVM.
- ***Zone Distribution Area (ZDA):*** ponto de interconexão opcional do cabeamento horizontal. Provê flexibilidade para o DATACENTER. Fica entre o HDA e o EDA.
- ***Equipment Distribution Area (EDA):*** área para equipamentos terminais (servidores, *storage*, unidades de fita) e os equipamentos de rede. Inclui RACKS e gabinete.

A **Figura 5-6** ilustra os componentes do modelo de DATACENTER em camadas:

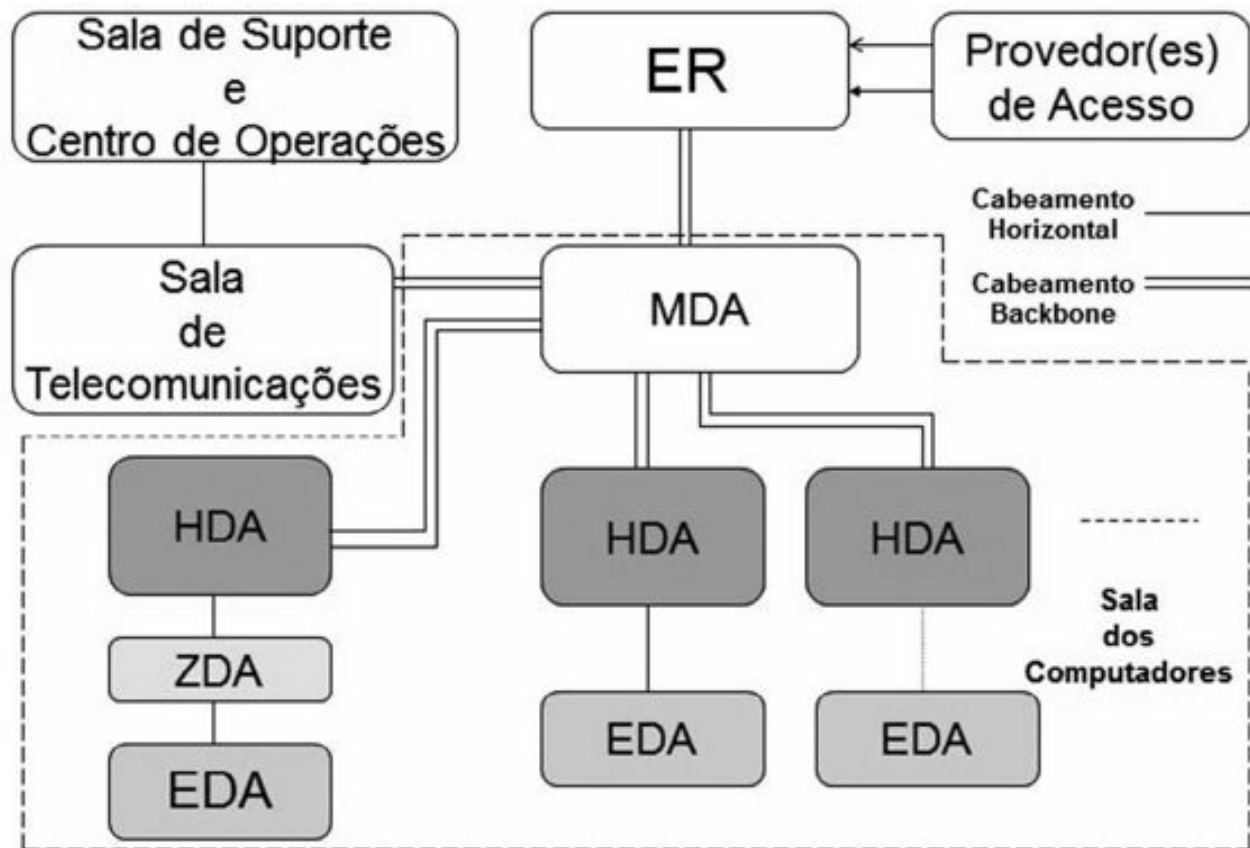


Figura 5-6 – Topologia básica de um DATACENTER segundo a TIA 942

Os servidores físicos são colocados em RACKS e a sua interligação feita por switches de acesso que se interligam a switches de agregação através de *uplinks* de 10 Gbps como padrão. Os switches de acesso em geral são do tipo ToR (*Top of Rack*) e os switches de agregação são do tipo EoR (*End of Row*).

5.2.3. Camadas (TIERS)

A norma TIA 942¹⁰ define a classificação dos DATACENTERS em quatro níveis independentes, chamados de TIERS (camadas), considerando a Arquitetura e Estrutura, Telecomunicações, Aspectos Elétricos e Mecânicos. Ou seja, para ter o DATACENTER em uma destas camadas ele deve obedecer aos requisitos nestas quatro frentes. O requisito da camada deve ser obedecido para Telecomunicações, Arquitetura e Infraestrutura, Telecomunicações, Sistemas elétrico e mecânico.

Tier 1: Básico

Neste modelo não existe redundância nas rotas físicas e lógicas do modelo. Prevê a distribuição de energia elétrica para atender a carga sem redundância. A falha elétrica pode causar interrupção parcial ou total das operações.

- **Pontos de falha:** falta de energia da concessionária no DATACENTER ou na Central de Operadora de Telecomunicações. Falha de equipamentos da Operadora. Falha nos

roteadores ou switches quando não redundantes. Qualquer catástrofe nos caminhos de interligação ou nas áreas ER, MDA, HDA, ZDA e EDA.

- **Downtime:** 28.8 hr/ano (equivale a 99.671%).

Tier 2: Componentes Redundantes

Os equipamentos de telecomunicações do DATACENTER e também os equipamentos de telecomunicações da operadora, assim como os dispositivos da LAN-SAN, devem ter módulos de energia redundantes. O cabeamento do *backbone* principal LAN e SAN das áreas de distribuição horizontal para os switches do *backbone* deve ter cabos de cobre ou fibras redundantes. Na TIER 2 deve-se ter duas caixas de acesso de telecomunicações e dois caminhos de entrada até o ER. Deve-se prover módulos de *no-break* redundantes em N+1. É necessária um sistema de gerador elétrico para suprir toda a carga. Não é necessário redundância na entrada do serviço de distribuição de energia. Os sistemas de ar condicionado devem ser projetados para operação contínua 24x7 e incorporam redundância N+1.

- **Pontos de Falha:** falhas nos sistemas de ar condicionado ou de energia podem ocasionar falhas em todos os outros componentes do DATACENTER.
- **Downtime:** 22 hr/ano (equivale a 99.749%).

Tier 3: Manutenção sem Paradas

Deve ser atendido por no mínimo duas operadoras de telecomunicações com cabeamentos distintos. Deve-se ter duas salas de entrada (ER) com no mínimo 20 metros de separação. Estas salas não podem compartilhar equipamentos de telecomunicações e devem estar em zonas de proteção contra incêndios, sistemas de energia e ar condicionado distintos. Deve-se prover caminhos redundantes entre as salas de entrada (ER), as salas MDA e as salas HDA. Nestas conexões deve-se ter fibras ou pares de fios de cobre redundantes. Deve-se ter uma solução de redundância para os elementos ativos considerados críticos como o *storage*. Deve-se prover pelo menos uma redundância elétrica N+1.

- **Pontos de falha:** qualquer catástrofe no MDA ou HDA irá interromper os serviços.
- **Downtime:** 1.6 hr/ano (equivale a 99.982%).

Tier 4: Tolerante a Falhas

Todo o cabeamento deve ser redundante, além de protegido por caminhos fechados. Os dispositivos ativos devem ser redundantes e devem ter pelo menos duas alimentações de energia ativas e independentes. O sistema deve prover comutação automática para os dispositivos de backup. Recomenda-se uma MDA secundária que, em zonas de proteção contra incêndio, devem ser separadas e alimentadas por caminho separado. Não é necessário um caminho duplo até o EDA. Deve-se prover disponibilidade elétrica 2(N+1). O prédio deve ter pelo menos duas alimentações de energia de empresas públicas a partir de diferentes subestações. O sistema de ar condicionado deve incluir múltiplas unidades com a capacidade de resfriamento combinada para manter a temperatura e a umidade relativa de áreas críticas nas condições projetadas.

- **Pontos de falha:** se a MDA primária falhar e não houver MDA secundária. Se a HDA primária falhar e não houver HDA secundária.
- **Downtime:** 0.4 hr/ano (equivale a 99.995%).

As áreas de requisitos para camadas da Norma TIA-942 são mostradas na [Figura 5-7](#).

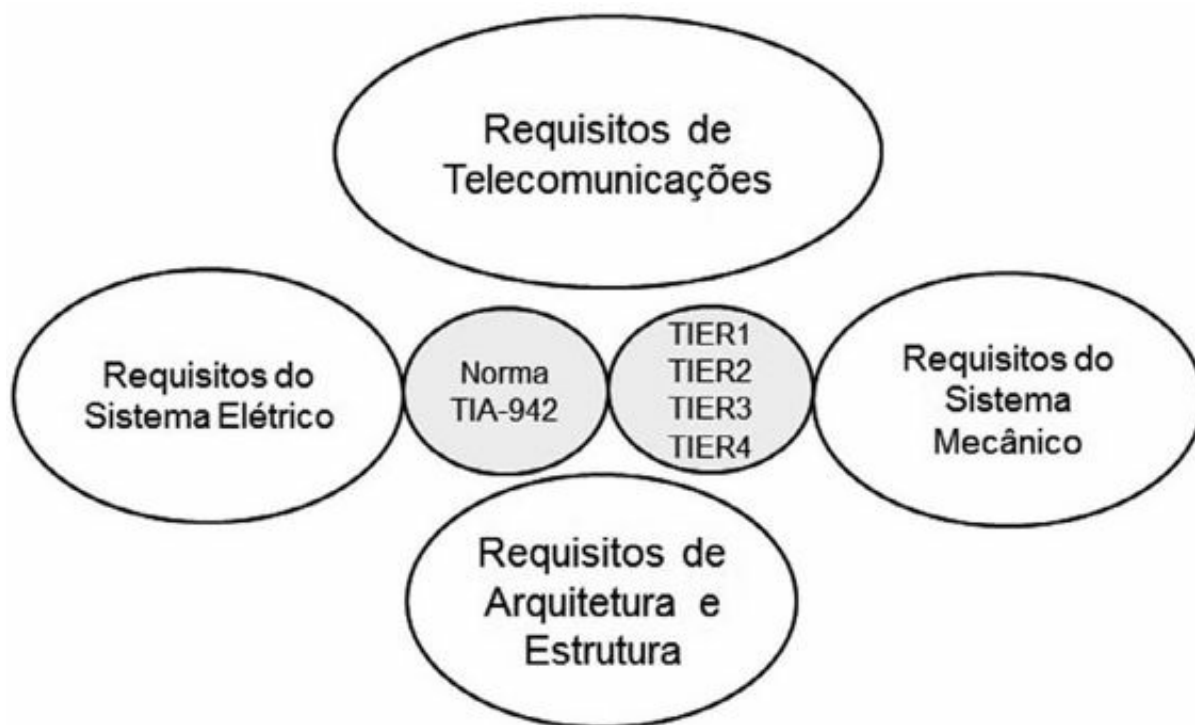


Figura 5-7 – Áreas de Requisitos para Camadas da Norma TIA-942

A [Tabela 5-2](#) resume os aspectos a serem considerados em um projeto para cada uma das áreas envolvidas.

Tabela 5-2 – Áreas e Aspectos a serem considerados da Norma TIA-9

Áreas	Aspectos
Telecomunicações	Cabeamento, RACKS, pathways, entrada de provedores, backbones, serviços de redundância, sala de entrada secundária, sala de distribuição secundária, roteadores e switches redundantes, patchpanels, patchcords, etc.
Arquitetura e Estrutura	Seleção do site Estacionamento Ocupação dentro do prédio Construção

Lobby de entrada
 Escritórios
 Escritório de segurança
 Centro de operação
 Áreas de apoio
 Salas de bateria e *no-breaks*
 Corredores
 Áreas de *shipping* e recebimento
 Áreas de gerador e combustíveis
 Segurança
 Segurança e Monitoração do Acesso
 Paredes, portas, janelas resistentes
 Monitoração CCTV
 CCTV
 Estrutural

Elétrica

Geral *No-breaks*
 Aterramento
 Sistema de desativação de sala de computadores
 Sistema de desativação de Bateria
 Sistema de desativação de Emergência
 Sistema de monitoração
 Configuração de bateria
 Baterias tipo *Flooded*
 Sala de Baterias
Enclosures para rotação de *no-breaks*
 Sistema de geração *Stand-by*
 Loadbank para teste
 Manutenção dos equipamentos

Mecânica

Geral
 Sistemas refrigerados a água
 Sistemas refrigerados com *Chiller*
 Rejeição de calor
 Sistemas refrigerados a Ar
 Sistemas de controle HVAC
Plumbing (para rejeição de calor em refrigeração a água)
 Sistemas de óleo combustível
 Supressão de fogo

A **Tabela 5-3** resume o *downtime* sugerido para cada camada da norma TIA 942 com um indicativo de custo (\$/W) de construção (só instalações sem a TI) nos Estados Unidos. Sair de um

modelo TIER 1 para TIER 3 em um DATACENTER de 1000KW significa sair de um projeto de US\$ 10M para um projeto de US\$ 20M. O DATACENTER padrão dos grandes provedores é o TIER 2, mesmo fora do Brasil. Também não adianta o DATACENTER ter um elevado nível de disponibilidade (TIER 3/4) se os equipamentos de TI apresentam baixo nível de disponibilidade.

Projetos atuais de grandes DATACENTERS (estado da arte) conseguem estar entre TIER 3 e TIER 4, mas a um custo muito menor (6\$/W) devido à otimização do projeto e à utilização de novas tecnologias.

Para saber qual a classificação de um DATACENTER perante o UPTIME basta consultar o link: <http://professionalservices.uptimeinstitute.com/tiercert.htm>

Tabela 5-3 – TIA 942 – *Downtime, Uptime* e Custo por Camada

	<i>Ano de Surgimento</i>	<i>Downtime (hr/ano)</i>	<i>Uptime (%)</i>	<i>Custo (\$/W)</i>
TIER 1	1965	28.8	99.671	10
TIER 2	1970	22	99.749	11
TIER 3	1985	1.6	99.982	20
TIER 4	1995	0.4	99.995	22
ESTADO DA ARTE (FORA DA NORMA)	2010	0.4 a 1.6	99.982 a 99.995	6

A Amazon divulgou recentemente que construiu seu DATACENTER por US\$ 88 milhões, para um consumo de 8MW de potência. Ressalta-se que este custo é só das instalações e não inclui os equipamentos de TI.

A HP¹¹ recentemente propôs o modelo de DATACENTER híbrido, um DATACENTER construído com diferentes níveis de criticidade, de forma a otimizar a relação entre custo e *downtime*. A **Figura 5-8** ilustra a ideia central. Construí-lo assim não é uma tarefa trivial.



Figura 5-8 – DATACENTER Híbrido

5.2.4. Certificações TIER

As certificações são fornecidas pelo UPTIME para os DATACENTERS de organizações em três modalidades:

- Documentos de projeto.
- Instalações construídas.
- Sustentabilidade operacional (ouro, prata ou bronze).

Recentemente, o Uptime Institute lançou uma certificação para indivíduos denominada *Accredited Tier Specialist* (ATS).

O DATACENTER da Ativas, em Belo Horizonte, MG, é certificado em projeto como TIER 3. Custou cerca de R\$ 50 milhões de dólares e está operacional desde o segundo semestre de 2010. Possui um tamanho de 6000 m² (1500 m² para área de TI). A potência elétrica pode chegar a 7,5 MW divididos em três etapas de crescimento de 2,5 MW. O DATACENTER da Ativas foi o primeiro a obter a certificação TIER 3 em documentos de projeto da América Latina.

A **Figura 5-9** mostra o DATACENTER das Ativas.

No Brasil existem atualmente cinco DATACENTERS certificados na categoria *Documentos de Projeto* como TIER 3: Ativas (MG), T-Systems (SP), Vivo (SP), Petrobras (RJ) e ALOG (SP).



Figura 5-9 – DATACENTER ATIVAS

5.3. Custo do DATACENTER

Não existe uma metodologia padrão para cálculo do custo total de propriedade (TCO) do DATACENTER. De uma forma geral, pode-se afirmar que o TCO de um DATACENTER é dado pela seguinte fórmula:

$$TCO (DC) = DATACENTER \text{ depreciação } DATACENTER \text{ Opex } TI \text{ depreciação } TI \text{ Opex}.$$

Na fórmula os custos do DATACENTER relacionados à construção das instalações, equipamentos de energia e refrigeração (DATACENTER depreciação e DATACENTER Opex) estão separados dos custos da carga de TI (TI depreciação e TI Opex) que envolve servidores, *storage*, redes e softwares.

O custo de capital (Capex) do DATACENTER varia largamente e depende de localização, projeto, velocidade da construção, classificação em TIER, etc.

Grandes DATACENTERS custam para construir em média \$10-\$22/W nos EUA, dependendo da camada de disponibilidade, conforme visto anteriormente. Novos DATACENTERS apresentam custos menores, da ordem de \$6/W. A depreciação utilizada em DATACENTER é tipicamente de 10 – 15 anos.

Os *determinantes* dos custos de construção do DATACENTER são:

- Energia, capacidade de refrigeração ou densidade (KW total, KW/cabinet ou W/ m²).
- TIER ou funcionalidade.
- Tamanho da sala de computadores (m² ou número de *cabinets*).

Os dois elementos primários do custo de capital são, assim, o componente KW, que varia em função do nível de disponibilidade (ver **Tabela 5-3**), e o componente “sala de computadores”. Em

todos os casos, \$2880/m² devem ser adicionados ao custo do KW. O terceiro elemento é o custo do espaço vazio construído para uso futuro. O valor a ser adicionado é \$1824/m². Valores monetários nos Estados Unidos.

O custo operacional (Opex) também é muito difícil de ser calculado, pois depende dos padrões utilizados. Nos Estados Unidos este custo, sem o custo da energia, é da ordem de \$0,02W-\$0,08W por mês. O custo operacional dos equipamentos de TI também varia muito e depende dos contratos acordados.

Exemplos de custos para um DATACENTER de 1900 m² são descritos abaixo:

- TIER II com 666 RACKS com 1.5 KW/rack (1 KW para computadores) = \$18,5 milhões.
- TIER IV com 666 RACKS com 3.0 KW/rack (2 KW para computadores) = \$56 milhões.

A conclusão deste estudo baseado nos dados do *Uptime Institute* é que os *determinante* primários para o custo de construção de um DATACENTER são as capacidades de energia e refrigeração e o tamanho da sala de computadores.

Estudos realizados pela APC sugerem custos de operação (opex) e custos de capital (capex) equivalentes para DATACENTERS em um típico ciclo de vida. Segundo a APC, a maior fonte de melhoria de TCO do DATACENTER é fazer o correto dimensionamento dos equipamentos de energia e refrigeração. O conceito fundamental é o de modularidade, a ser explicado mais adiante. A energia pode representar de 20% a 40% do TCO de um DATACENTER.

Um DATACENTER é um ativo que concentra grande parte do investimento de TI (25% em média para grandes organizações) e determinar o seu real TCO (*TrueTCO*) é uma tarefa importante. Jonathan Koomey, do UPTIME Institute, no artigo *A Simple Model for Determining True TCO for Datacenters*, sugere que os esforços para medir o TCO feitos até agora são importantes, mas são incompletos e pouco documentados. Koomey sugere uma forma simples de medir os custos e utiliza um DATACENTER para aplicações do tipo HPC como modelo. Os custos de capital das instalações se mostraram maiores do que os custos da TI.

5.4. Padronização do DATACENTER

Os principais *form-factors* utilizados em DATACENTERS são o CHASSI, o RACK e o CONTAINER, descritos a seguir.

5.4.1. CHASSI

Possibilita construir os DATACENTERS baseados em blocos do tipo CHASSIS que utilizam servidores em lâminas (*blades*), conforme ilustra a **Figura 5-10**.



Figura 5-10 – *Chassis Blade* da IBM

5.4.2. RACK

Possibilita construir os DATACENTERS da forma convencional em blocos de RACKS de servidores, conforme ilustra a **Figura 5-11**, permitindo obter uma modularidade média sem o compromisso de utilizar chassis de *blades* para o *form-factor* de todos os servidores. Os RACKS são construídos em três tamanhos principais: 24U, 42U e 48U, onde $1U = 44,45\text{mm}$.



Figura 5-11 – RACKS padrões 24U e 42U da DELL

A **Figura 5-12** ilustra as dimensões do RACK DELL:

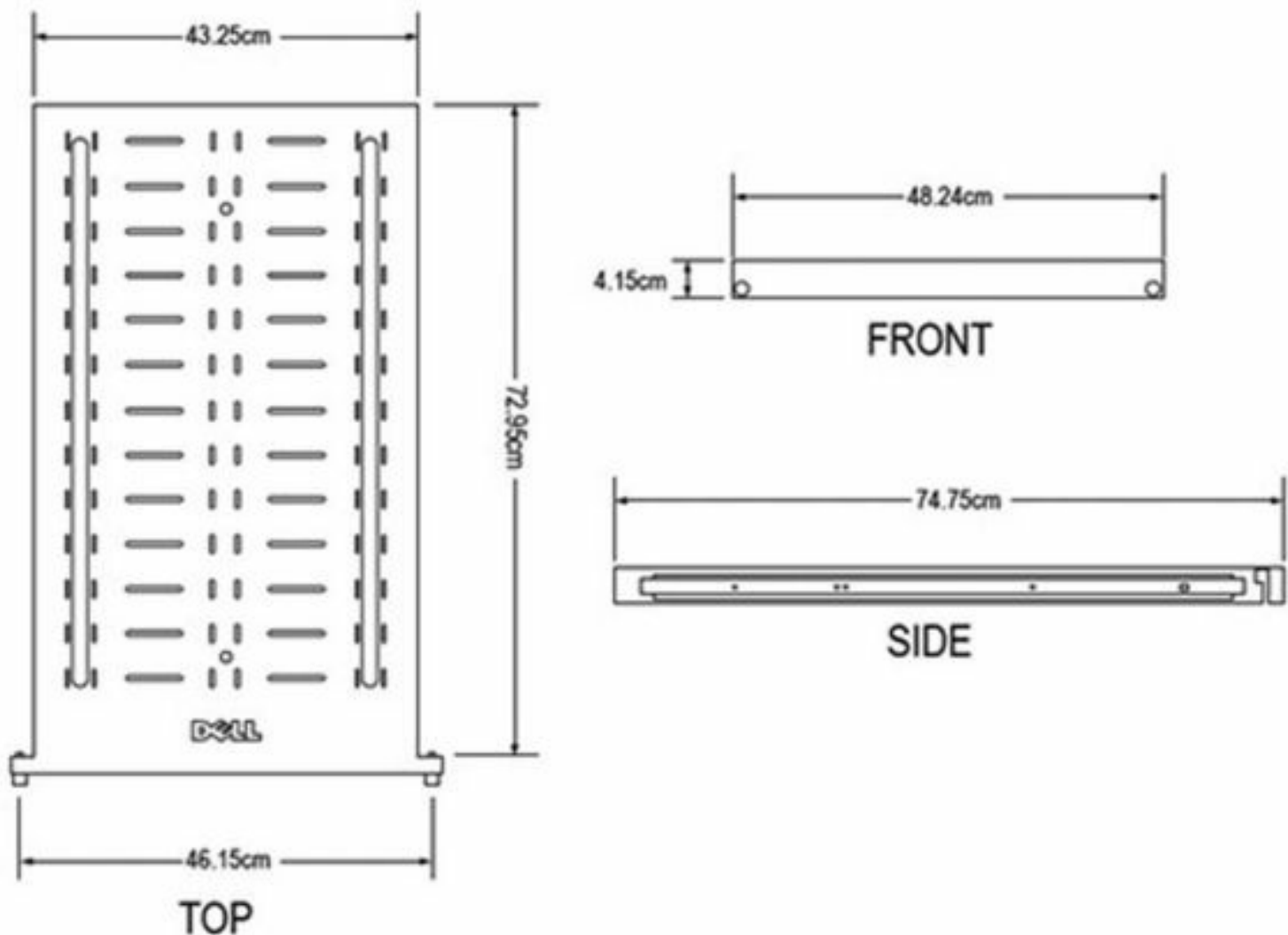


Figura 5-12 – Dimensões do RACK Dell

Os RACKS também podem ser construídos sob encomenda.

Recentemente o Facebook disponibilizou o documento OPEN COMPUTE PROJECT, que mostra como o projeto do seu novo DATACENTER foi pensado e projetado. A **Figura 5-13** ilustra o perfil dos RACKS utilizados. São na verdade TRIPLETS (3 RACKS em conjunto). O documento utilizado como referência é o *Server Chassis and Triplet Hardware 1.0*, disponibilizado no site <http://opencompute.org/datacenters/>.

No triplet RACK, os três RACKS são de 42U e cada um dos RACKS pode conter 30 servidores. Um gabinete de bateria para backup é colocado para cada dois conjuntos de três RACKS. A bateria fornece corrente contínua (direct current – DC), em caso de falta de alimentação da corrente alternada (alternate current – AC) dos RACKS.

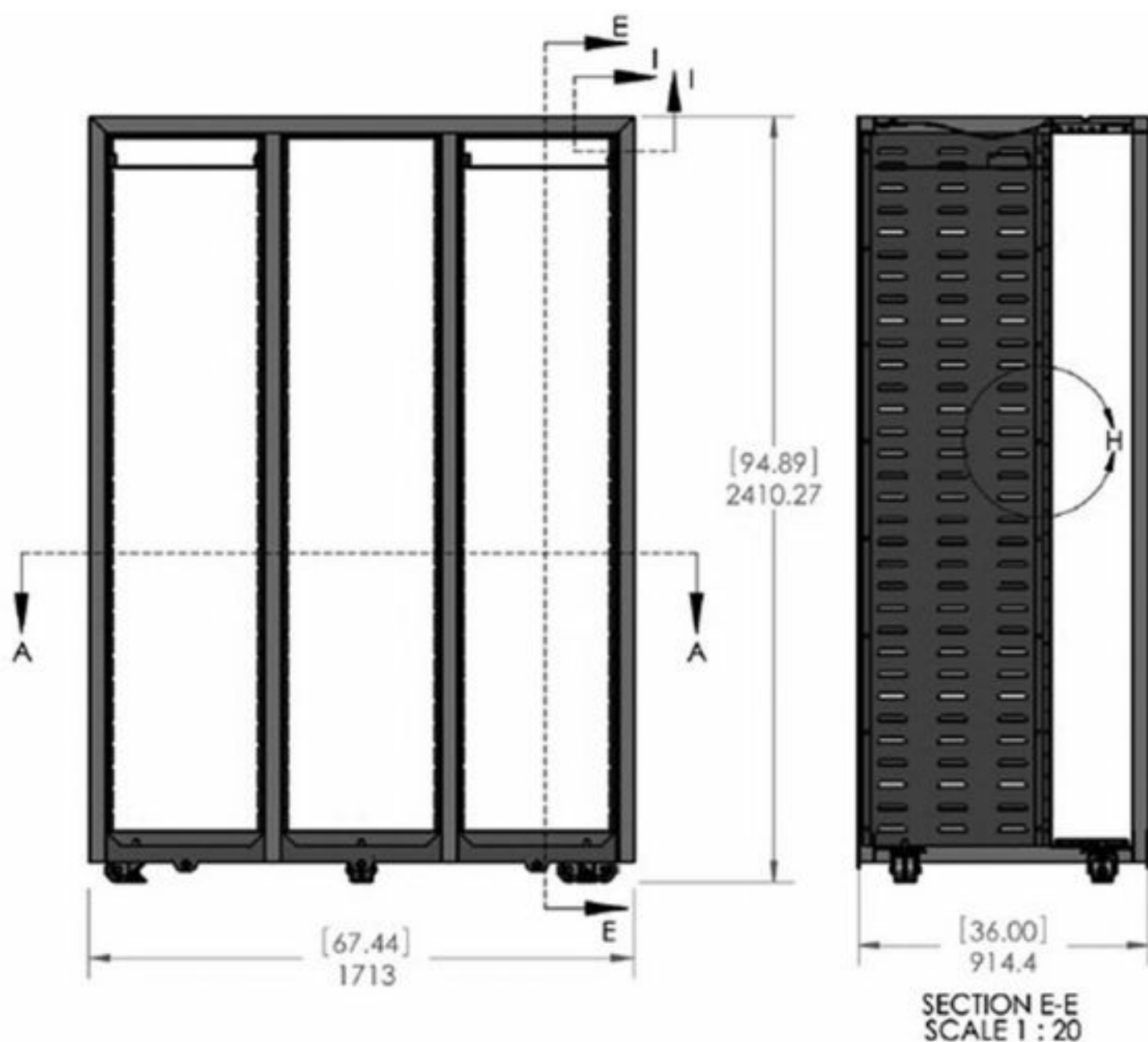


Figura 5-13 – Open Compute Project Triplet Dimensions

5.4.3. CONTAINER

Os RACKS de servidores podem ser embutidos em CONTAINERS que permitem o fácil deslocamento entre prédios, facilitando a montagem e permitindo a escalabilidade de equipamentos de energia, processamento e armazenagem.

Empresas como a ACTIVEPOWER estão portando os equipamentos de energia (*no-breaks* e geradores) e de resfriamento de água (*chillers*) em CONTAINERS específicos e estão estabelecendo parcerias com empresas como HP e Oracle (Sun). A solução PowerHouse, por exemplo, fornecida pela ACTIVEPOWER, é integrada com a solução CONTAINER da Sun, o MD-20, que permite ter RACKS internos com até 25KW de potência. Neste caso, seriam necessários dois CONTAINERS, um para TI e outro para energia e refrigeração, para compor a solução total.

A solução PowerHouse utiliza *no-breaks* baseados na tecnologia *flywheel* que dispensam o

uso de baterias e podem alimentar CONTAINERS de até 1MW. Os DATACENTERS baseados em CONTAINERS normalmente utilizam entre 1000 e 2000 servidores formando um POD (*Performance Optimized Datacenter*) que só necessita de energia, refrigeração e telecomunicações.

Os CONTAINERS são interessantes para DATACENTERS que podem ser disponibilizados rapidamente em eventos, para sites de contingência, em megainstalações que precisam ser escaláveis e até como DATACENTERS principais. Empresas como HP, IBM, Sun e Cisco já fornecem estas estruturas e Microsoft e Google atualmente utilizam CONTAINERS próprios.

A **Figura 5-14** ilustra o CONTAINER da Cisco. A principal vantagem destas estruturas é possibilitar o aumento da eficiência energética, basicamente melhorar a relação entre a energia que é adquirida da concessionária e a energia que efetivamente é consumida pela carga de TI ao longo do ciclo de vida do DATACENTER.

No caso do CONTAINER Cisco, ele possui 12,2 metros de comprimento e 2,4 metros de largura por 2,9 metros de altura. Pode suportar uma carga de 400KW e deve ser alimentado por uma tensão AC de 400 V trifásico. A máxima potência por RACK é 25KW, e as estruturas internas de TI são do tipo vBlock, a serem vistas mais adiante. São 16 RACKS de 44RU por CONTAINER, o que totaliza 704RU. O CONTAINER Cisco é refrigerado a água.



5.5. Instalação e Construção do DATACENTER

5.5.1. Introdução

O local de instalação do DATACENTER deve considerar uma seleção apropriada do local incluindo provimento de energia, telecomunicações, água, gás, qualidade do ar e influências geográficas como a possibilidade de ocorrência de tsunamis, terremotos, etc.

A construção do DATACENTER envolve calcular a necessidade de energia e refrigeração, cabeamento, controles de umidade e temperatura, sistemas de prevenção contra incêndios; segurança física, espaço para os *RACKS* e muitas vezes piso elevado. A norma TIA-942 ajuda nesta tarefa.

Energia e Refrigeração, Temperatura e Umidade são aspectos críticos a serem considerados no projeto de DATACENTERS.

- **Energia e Refrigeração:** o provimento de energia e refrigeração deve estar de acordo com as premissas adotadas quando da formulação do projeto. A carga de TI e a carga dos equipamentos que suportam a refrigeração e o seu crescimento futuro devem ser considerados no correto dimensionamento dos componentes chaves do DATACENTER.
- **Temperatura e Umidade:** devem ser monitoradas e controladas por meio de sistemas de climatização precisos e redundantes. Os DATACENTERS utilizam em muitos casos pisos elevados para possibilitar que o fluxo de ar seja insuflado para baixo dos equipamentos, a fim de mantê-los refrigerados.

5.5.2. Modularidade

Modularidade é o aspecto fundamental para a obtenção da eficiência energética. Eficiência energética do DATACENTER é tema do próximo capítulo.

O que seria a Modularidade?

- Modularidade é o uso de unidades separadas, mas funcionais, que podem trabalhar juntas e que possibilitam uma forma mais efetiva para crescimento do DATACENTER.
- Modularidade é utilizar para crescer incrementos menores de componentes padrões que possibilitam casar os requisitos TI com os requisitos de negócio e adicionar capacidade quando necessário.

A modularidade:

- Permite ter melhor escalabilidade no DATACENTER;
- Pode ser conseguida com DATACENTERS tradicionais (TDC) ou com DATACENTERS em CONTAINERS (CDC);
- Fundamentalmente busca o aumento da eficiência energética e a redução do espaço físico;

- Permite que a Infraestrutura (equipamentos de energia e refrigeração) cresça de acordo com o crescimento da demanda (carga de TI).

5.6. Visão Geral dos DATACENTERS em CONTAINERS

5.6.1. Introdução

A modularidade através da utilização de CONTAINERS iniciou-se em 2007 com a oferta da Sun chamada de Project BlackBox. O Google revelou posteriormente que já tinha construído um DATACENTER em CONTAINER em 2005 após desenvolver o conceito em 2003. Atualmente os DATACENTERS em CONTAINERS são ofertados por diversos fabricantes e representam uma opção aos DATACENTERS tradicionais. Agora esta opção é oferecida por diversos fornecedores (os tamanhos normalmente são de 20 e 40 pés, mas podem ser feitos sob encomenda com alguns fabricantes como a I/O ANYWHERE). As principais opções disponíveis hoje são:

- **CISCO:** Cisco CDC
- **DELL:** HUMIDOR
- **HP:** POD
- **I/O ANYWHWRE:** I/O
- **IBM:** PMDC
- **NxGEN Modular:** NxGEN
- **SGI:** ICE CUBE

No Brasil empresas como a ACECO TI e a GEMELO estão produzindo DATACENTERS em CONTAINERS já em 2011 em diversos formatos. A **Figura 5-15** mostra o DATACENTER em CONTAINER da I/O ANYWHERE.



Figura 5-15 – DATACENTER CONTAINER da I/O ANYWHERE.

Google e Microsoft também desenvolveram seus próprios CONTAINERS para grandes DATACENTERS com foco na melhoria da modularidade e no tempo de construção de novos DATACENTERS. Os CONTAINERS também podem ser utilizados em pequenas instalações como DATACENTERS primários ou mesmo DATACENTER de contingência.

A adição de capacidade no formato de CONTAINERS é facilitada. A melhor eficiência energética inerente ao CONTAINER resulta em custos operacionais mais baixos. Com os CONTAINERS existem as dificuldades com segurança física, fornecimento de energia e refrigeração e geração de energia back-up.

As primeiras ofertas de CONTAINERS propiciavam poucas melhorias em termos de eficiência energética e utilizam refrigeração baseada em *chillers* internos ou externos ou mesmo faziam o resfriamento baseado em expansão direta (unidades *DX cooling coil*) *on-board*.

Segundo o relatório do LBNL¹², em regiões onde o clima não é frio, os DATACENTERS baseados em CONTAINERS de primeira geração podem continuar a ser uma boa opção. A segunda geração de CONTAINERS utiliza *air-side economizers* e não necessitam de *chillers* baseados em água, tornando a infraestrutura mais simples e reduzindo os custos.

O PUE (*Power Utilization Effectiveness*), tema central do próximo capítulo, que representa a relação entre a energia consumida pelo DATACENTER e a energia consumida pela carga de TI, pode chegar a 1.16 nos CONTAINERS de primeira geração e 1.02 nos CONTAINERS de segunda geração. O projeto, baseado em uma refrigeração *close-coupled*, comum em CONTAINERS de primeira geração, permite obter uma maior eficiência. A segunda geração utiliza *air-side economizers* que melhoram ainda mais o PUE. Cabe ressaltar que o PUE não pode ser menor que 1.

A ideia central da modularidade em CONTAINERS é que a adição de capacidade é facilitada por estas estruturas quando comparada a instalações convencionais, que normalmente envolvem obras civis, resultando em menores custos de capital e de operação. Obviamente, estas estruturas precisam de suporte adequado de infraestrutura de energia e refrigeração.

5.6.2. Seleção de CONTAINERS

O guia do LBNL, citado na referência deste capítulo, sugere um processo de seleção de CONTAINERS para DATACENTERS resumido a seguir:

■ Requisitos dos equipamentos de TI

- ☐ Espaço inicial e futuro para os RACKS.
- ☐ Energia inicial, média e futura de TI por RACK.
- ☐ Temperatura máxima interna para o equipamento de TI.
- ☐ Temperatura máxima e mínima de refrigeração de saída do ar.
- ☐ Tipo de fluxo de ar por equipamento de TI.
- ☐ Conexões de energia por RACK.

■ Seleção do tipo de refrigeração mais eficiente

- ☐ Opções são *chillers* refrigerados a água, *chillers* refrigerados a ar, refrigeração baseada em compressores DX.

■ Decisão sobre requisitos adicionais

- ☐ Segurança do site.
- ☐ Proximidade a energia e *chillers* baseados em água.
- ☐ Equipamentos de energia backup.
- ☐ Sistemas de supressão de fogo e detecção de fumaça.
- ☐ CONTAINERS com *air-side economizers* devem ter uma orientação correta em relação ao vento.

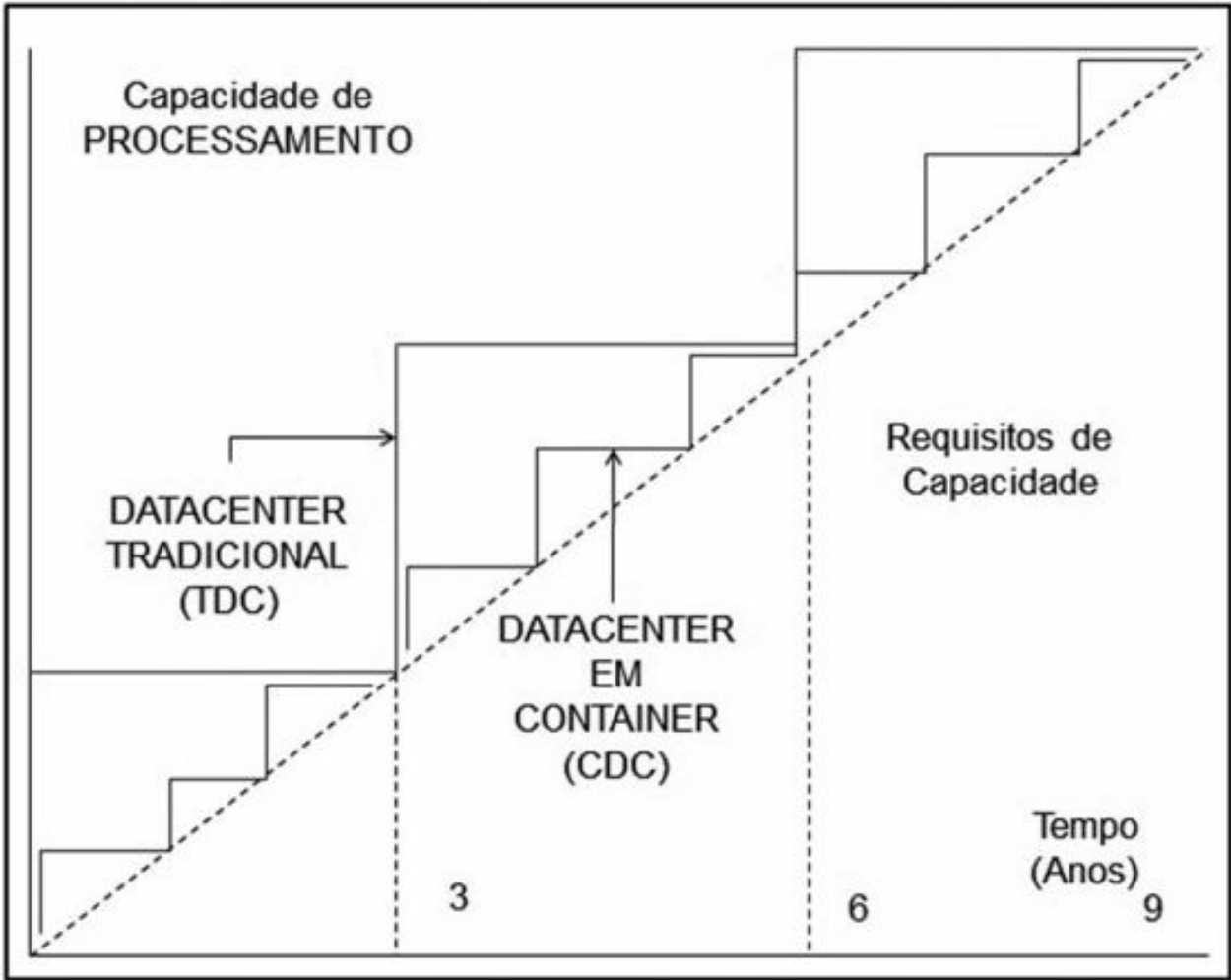
5.6.3 DATACENTER em CONTAINER (CDC) versus DATACENTER Tradicional (TDC)

A **Tabela 5-4**, baseada no modular Data Centers Procurement grille do LBNL, ilustra as diferenças entre as versões de DATACENTERS comparando os custos de capital e de operação e o tempo para desenvolver.

Tabela 5-4 – Comparação CDC com TDC

Atributos Primários	TDC	CDC Primeira Geração	CDC Segunda Geração
Tempo para Implementar	Longo – tipicamente dois anos.	Meses – potencialmente curto.	Meses – simplificado pela menor estrutura de refrigeração quando comparado ao CDC Primeira Geração.
Custo de Capital	Muito alto – US\$ 20M por MW.	Mais baixo – Falta informação sobre custos de implantação.	Mais baixo – Custo aumentado devido à redução dos custos de infraestrutura.
Custo de Operação	Variável com PUE de 2.0 até 1.2.	Similar ao TDC com alguns ganhos e pré-engenharia.	Similar ao melhor TDC que usa refrigeração a ar.

A grande questão que se coloca é que os DATACENTERS em CONTAINERS podem ser mais rapidamente implementados do que os DATACENTER convencionais, pois o projeto não envolve obras de engenharia diretamente. Também a sua escalabilidade é facilitada, pois é feita em incrementos menores, conforme ilustra a [Figura 5-16](#).



A **Figura 5-17** ilustra a diferença na forma de fazer o crescimento de DATACENTERS em CONTAINERS (a) contra a forma tradicional (b). A ideia anterior, muito utilizada, de fazer um prédio de grande capacidade e depois crescer, normalmente implica em baixa eficiência.

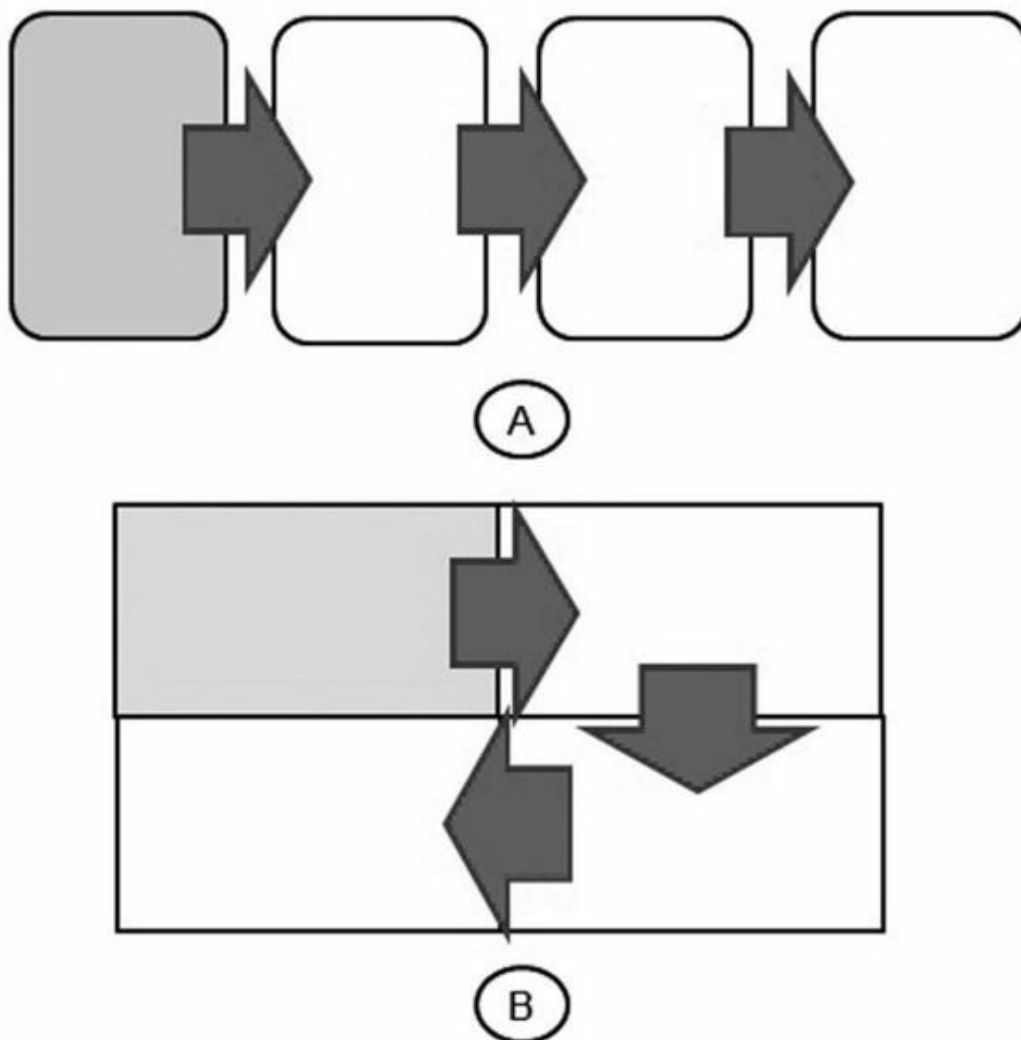


Figura 5-17 – Crescimento dos DATACENTERS (A) Baseado em CONTAINERS (B) Tradicional

5.7. Segurança Física do DATACENTER

A segurança física é um aspecto crucial da segurança do DATACENTER. O acesso do pessoal no DATACENTER deve ser controlado. Minimizar a quantidade de pessoas que podem entrar no DATACENTER e assegurar o controle é uma medida para mitigar o risco. A segurança física pode ser pensada em termos de profundidade de segurança (*depth of security*), ou seja, é necessário pensar a segurança em diferentes níveis partindo de níveis mais leves de segurança, no estacionamento, por exemplo, para níveis mais fortes de segurança próximos à carga de TI. A **Figura 5-18** ilustra a segurança de perímetro no DATACENTER.

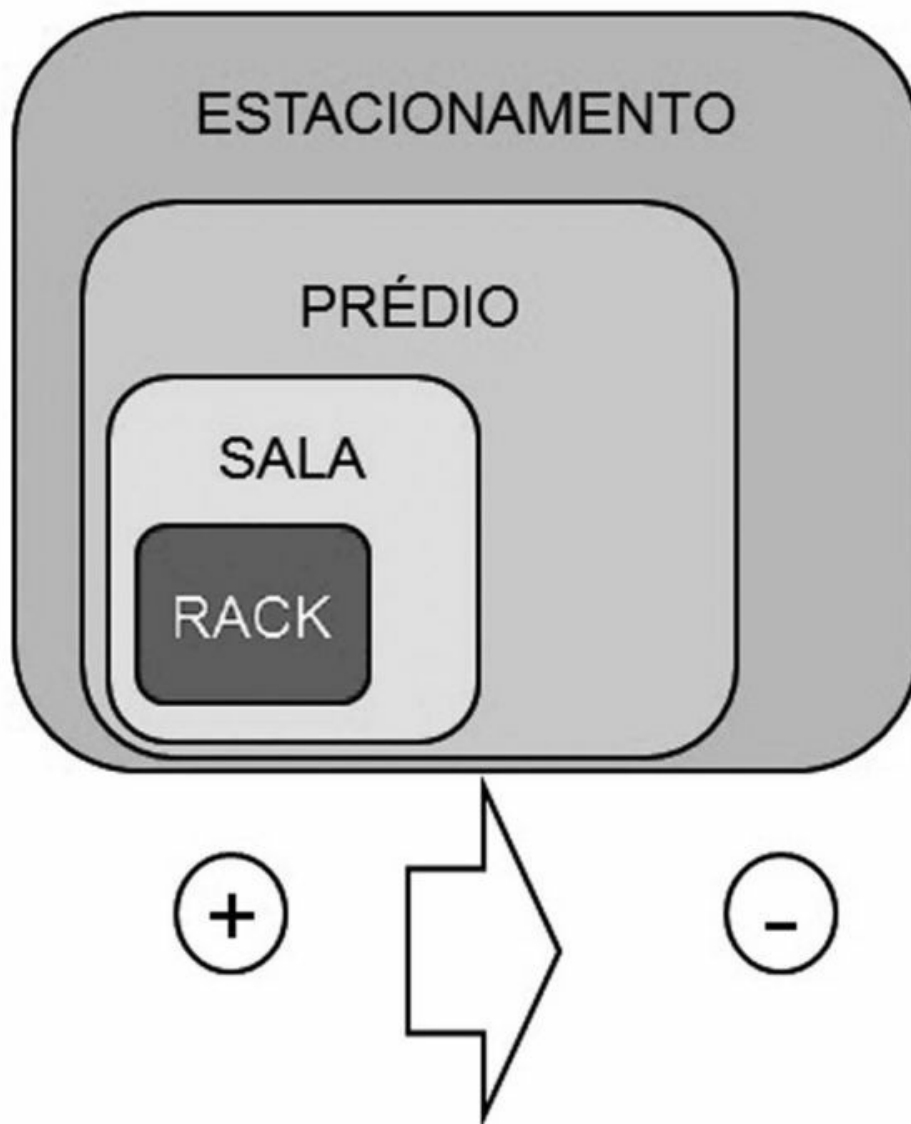


Figura 5-18 – Segurança Física do DATACENTER

Também a recuperação de desastres deve ser pensada. Em bancos é comum a existência de até três DATACENTERS CORPORATIVOS em distâncias distintas que permitem a recuperação de desastres quase perfeita.

Segundo o instituto SANS, no documento *SANS Institute Reading Room*, aspectos fundamentais a serem considerados em termos de segurança física são:

■ **Localização do site**

- ☐ Riscos de desastres naturais.
- ☐ Riscos de desastres causados pelo homem.
- ☐ Infraestrutura de energia e telecomunicações adequada.
- ☐ Estrutura do DATACENTER ter propósito único.

■ **Perímetro do site**

- ☐ Vigilância.

- Local das salas de computadores.
- Pontos de acesso.

■ Salas de computadores

- Acesso.
- Infraestrutura.
- Ambiente: temperatura (55 a 75 F) e umidade (20% a 80%).
- Prevenção de incêndio.
- Espaço compartilhado.

■ Instalações

- Torres de refrigeração redundantes.
- Conjunto de baterias deve permitir o chaveamento para o gerador.
- Documentos devem ser destruídos sempre que necessário.
- O Network Operation Center (NOC) deve ter sistemas de monitoração de energia, clima, temperatura e umidade. Deve operar 24 horas por dia e ter meios redundantes de comunicação.

■ Recuperação de desastres

- **Plano de recuperação:** deve estabelecer o que constitui um desastre? Quem deve ser notificado? Qual o procedimento? Em que condições o plano deve ser atualizado?
- **Backup offsite:** deve ser feito regularmente.
- **Site redundante:** servidores redundantes devem estar em outro site.

5.8. Gerenciamento do DATACENTER

O gerenciamento e a monitoração de todo o ambiente também são aspectos relevantes. O *Network Operation Center* (NOC) é a célula central do gerenciamento. Por definição, o NOC é um local onde se centraliza a gerência da rede e dos componentes do DATACENTER. A partir desse centro e de aplicações que monitoram a rede os administradores podem saber, em tempo real, a situação de cada “componente” dentro da rede. Os componentes são os servidores, roteadores, switches, *storage*, etc.

O NOC normalmente é utilizado em grandes DATACENTERS, mas a tendência é que as pequenas e médias empresas também adotem este tipo de gerenciamento. O NOC pode ser inclusive terceirizado com um provedor que dilui o seu custo com a venda do serviço de NOC para outras empresas.

5.8.1. DCIM

A proposta do *Datacenter Infrastructure Management* (DCIM) é fazer o gerenciamento de todo o DATACENTER integrando infraestrutura e TI. DCIM mantém uma fonte com detalhes técnicos de consumo de energia e rede para os dispositivos de TI envolvidos na topologia do DATACENTER, otimiza os recursos utilizados na energização e no resfriamento do DATACENTER e pode até ser utilizado no planejamento da capacidade quando integrado a outras ferramentas de software.

O estágio atual da monitoração e automação de DATACENTERS é de ferramentas de empresas de TI como BMC, CA, HP, IBM, Microsoft e VMware e companhias de infraestrutura como Emerson e Schneider e o surgimento de ferramentas que permitem o completo gerenciamento do DATACENTER.

A proposta do DCIM é ser um único sistema para diagnosticar toda a infraestrutura de TI e as instalações. A **Figura 5-19** ilustra a utilização do DCIM.

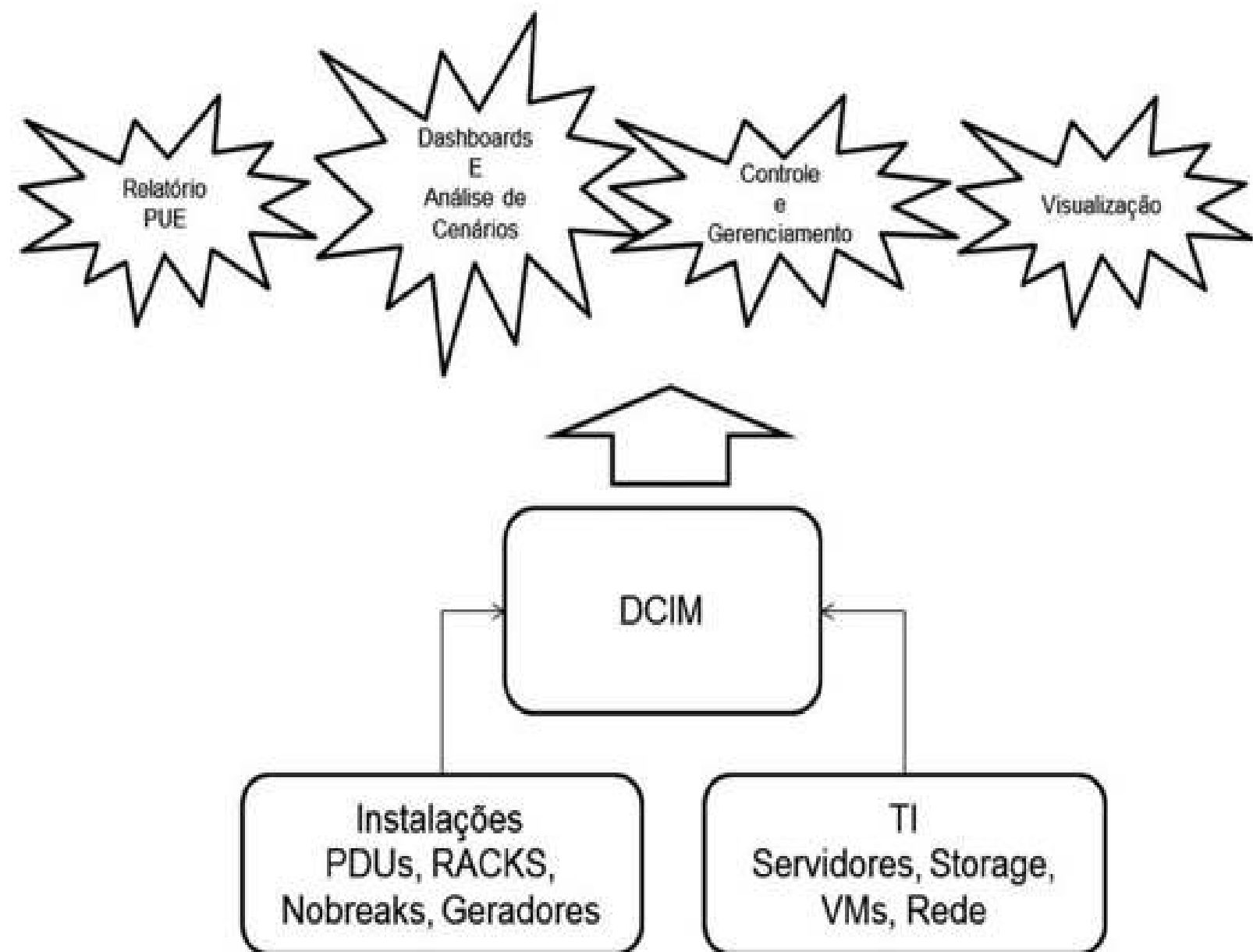


Figura 5-19 – Utilização do DCIM

5.9. Referências Bibliográficas

- A Simple Model for Determining True Total Cost of Ownership for Data Centers, Uptime Institute, 2007.
- An Introduction to Data Center Infrastructure Management, Raritan, 2010.
- Belady, Christian, Projecting Annual New Datacenter Construction Market Size, GFS, Microsoft, 2011.
- Cost Model: Dollars per KW plus Dollars per Square Foot of computer Floor. Uptime Institute, 2008.
- Cost Model for Planning, Development and Operation of a DATACENTER Chandrakant D. Patel, Amip J. Shah, Internet Systems and Storage Laboratory, HP Laboratories Palo Alto. HPL-2005-107(R.1), June 9, 2005.
- Guidelines for Specifying DATACENTER Critically/Tier Levels. APC White Paper #122.
- IT Consolidation: Business Drivers, Benefits and Vendor Selection, IDC Executive Brief, 2005.
- Luiz André, Barroso & Hölzle, Urs. The DATACENTER as a Computer. An Introduction to the Design of Warehouse-Scale Machines, 2009.
- Modular/CONTAINER DATA CENTERS Procurement Guide: Optimizing for Energy Efficiency and Quick Deployment. Lawrence Berkeley National Laboratory (LBNL), February 2011.
- Open Compute Project, Data Center v1.0, April, 2011.
- Open Compute Project, Server Chassis and Triplet Hardware v1.0, April 2011.
- Physical Security in Mission Critical Facilities. APC White Paper #82.
- Tier Classification Define Site Infrastructure Performance, Uptime Institute, 2008.
- Uma arquitetura aperfeiçoada para Centros de Dados de alta densidade e alta eficiência. APC White Paper #126.
- Veras, Manoel. DATACENTER: Componente Central da Infraestrutura de TI, Brasport, 2009.

6. DATACENTER: Eficiência Energética

6.1. Introdução

A busca pela melhoria da eficiência energética é o que há de mais atual em projeto de DATACENTERS, estrutura básica para a construção da nuvem. Nos últimos três anos verificou-se uma corrida de grandes organizações para melhorar a eficiência energética de novos projetos de DATACENTERS ou mesmo de DATACENTERS em operação. A razão desta corrida é principalmente a busca da redução dos custos de operação do DATACENTER e os requisitos de sustentabilidade.

Sustentabilidade é um conceito sistêmico relacionado à continuidade dos aspectos econômicos, sociais, culturais e ambientais da sociedade humana. Precisa-se considerar que os aspectos socioambientais exercem influência crescente nas decisões das empresas e que o desenvolvimento sustentável é a ideia de não esgotar os recursos para o futuro.

A energia utilizada para alimentação do DATACENTER é uma das fontes de preocupação do mundo sustentável. Como o consumo de energia por estas estruturas é muito grande, é necessário pensar em novas formas de alimentação, de preferência usando fontes renováveis de energia. Novos projetos têm contemplado estas novas fontes em uma solução híbrida para a alimentação dos DATACENTERS. Energia do vento (eólica), por exemplo, pode ser utilizada, mas a sua variabilidade faz quase que sempre que seja considerada parte de um *grid* energético e não uma solução isolada.

Boa parte dos DATACENTERS nos Estados Unidos, por exemplo, é alimentada por uma matriz energética baseada em energia nuclear e carvão. Os donos dos DATACENTERS estão sendo pressionados a mudarem este quadro acrescentando energias alternativas como a energia eólica e a energia solar.

A criação do Green Grid em 2007, organização formada por fabricantes da área de TI, foi um marco na busca de métricas que possibilitem aferir a eficiência energética dos DATACENTERS. Recentemente novas métricas relacionadas a consumo de água e emissão de carbono mostram que a corrida em busca de projetos sustentáveis continua. Viva a indústria de TI, que reage rapidamente às demandas da sociedade e das organizações.

6.2. Eficiência Energética do DATACENTER

Conforme mencionado anteriormente, é necessário considerar aspectos ambientais no projeto de novas estruturas e na operação da nuvem considerando o grande consumo de energia provocado por estas estruturas que são cada vez mais densas e, portanto, grandes geradoras de calor. O

DATACENTER é um elemento chave da infraestrutura e onde acontece boa parte do consumo de energia do ambiente de TI. Projetá-lo obedecendo a aspectos de maior eficiência energética é o aspecto relevante do momento.

Um aspecto relevante em projetos é sempre considerar a potência real consumida nos projetos para cálculo da carga de TI. Um servidor com 1600 W de fonte de alimentação não consome este valor. A plena carga consumirá em torno de 67% do valor nominal da fonte. Sem este ajuste pode-se incorrer no erro de projetar um DATACENTER pelo valor nominal da fonte, superdimensionando os equipamentos de infraestrutura.

Outra confusão que se faz é misturar o conceito de Potência (KW) e Energia (KWH). Energia é potência multiplicada por tempo. Se um servidor consome 5 KW em 1 hora, a energia consumida será de 5 KWH. Se consumir 5 KW em 30 minutos, a energia consumida será de 10 KWH. O consumo de energia é dado pela seguinte fórmula:

$$\text{Consumo de um Servidor (KWH)} = \text{Potência (KW)} \times \text{Número de Horas de Uso/dia} \times \text{Número de Dias de Uso/mês.}$$

Se um servidor consome 1.2 KW e é utilizado 30 minutos por dia, tem-se um consumo de 0.60KWH por dia. Se o servidor é utilizado 30 dias por mês, tem-se um consumo de 18KWH mensal.

O preço a ser pago seria:

$$\text{Valor do Consumo Ativo (R\$)} = \text{Consumo do Período em KWH} \times \text{Preço do KWH}$$

Neste caso, Valor (R\$) = $18 \times 0.34487 = 6,2$ (0,34487 = CONSUMO B3 INDUSTRIAL referência Brasil)

Também a diferença entre KVA e KW deve ser considerada. Esta diferença é traduzida pelo fator de potência, que é normalmente um número menor que 1, indicando que a potência em KW é quase sempre menor do que a potência em KVA. Fator de Potência = KW/KVA.

A eficiência do DATACENTER, até bem pouco tempo atrás, era medida unicamente em termos de indicadores vinculados à disponibilidade e ao desempenho. Com os aspectos ambientais sendo cada vez mais considerados, o aumento dos custos de energia e a limitação no fornecimento de energia por parte de alguns provedores de energia, é natural que os gerentes de infraestrutura de TI repensem a estratégia para o DATACENTER e considerem o aspecto da sustentabilidade nas diversas escolhas que precisam fazer, incluindo engenharia, equipamentos, tecnologias e a própria operação.

Estudos realizados na Universidade de Stanford indicam que o consumo de energia dos DATACENTERS representa 1,2% de todo o consumo de eletricidade nos EUA. O *Uptime Institute* levantou que os custos de energia representam hoje até 44% do TCO de um DATACENTER.

Com o aumento dos custos com energia, provocado por servidores cada vez mais densos, é

natural que a responsabilidade pelos gastos com energia do DATACENTER deve ir para dentro do departamento de TI, e neste momento sugiro que você já tenha adequado a sua estratégia verde ao DATACENTER.

A eficiência do DATACENTER em operação, construído há mais de quinze anos, em geral não passa de 40%. Ou seja, de 100% de energia que era injetada no DATACENTER só 40% alimentava a carga de TI. Em média, 60% da energia que alimentava o DATACENTER na verdade era consumida antes de chegar à carga de TI.

A **Figura 6-1** ilustra as fontes de consumo de energia no DATACENTER típico, baseado em referência da APC. A indústria está fazendo um esforço enorme para mudar este quadro e novos projetos de DATACENTER consideram novas formas de fornecimento de energia e otimização do tamanho dos equipamentos de energia e refrigeração como premissas. A forma de resfriar o DATACENTER é sem dúvida o grande gargalo que provoca boa parte da ineficiência do DATACENTER.

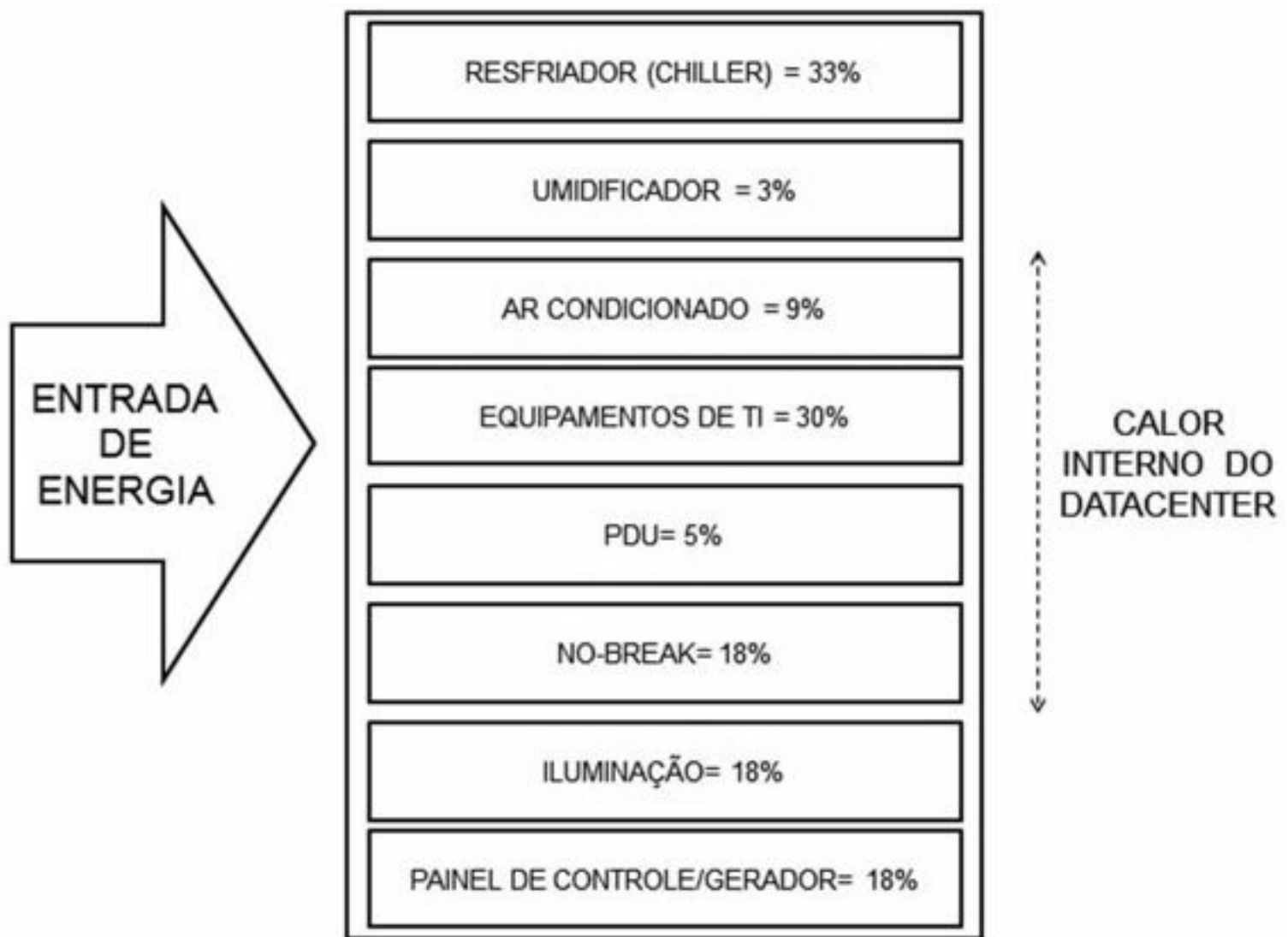


Figura 6-1 – Fontes de Consumo de Energia em um DATACENTER típico

O artigo *DATACENTER Efficiency in the Scalable Enterprise*, publicado na revista Dell Power Solutions em 2007, mostra que a Dell monitorou em laboratório o consumo de um

DATACENTER típico, construído há alguns anos, e verificou que, de 100% da energia injetada no DATACENTER, só 41% efetivamente foi utilizada pela carga de TI, 28% foi consumida pelos equipamentos que transformam e adequam a energia que chega à carga de TI e 31% foi consumida pelos equipamentos de suporte (refrigeração). Este artigo reforça a ideia de que a ineficiência é comum em projetos de DATACENTERS de alguns anos atrás. Na verdade, existe um descasamento provocado por instalações construídas em torno do máximo da capacidade e uma carga de TI que está muito longe de consumir a energia disponível.

Este *oversizing* (superdimensionamento) do DATACENTER chegava até três vezes a capacidade necessária e era uma prática corriqueira de alguns anos atrás que gerava custos de capital e manutenção em excesso. Este custo pode ser reduzido implementando uma arquitetura que possa se adaptar a requisitos de capacidade que mudam com o tempo.

A APC sugere, como boa prática, a construção do DATACENTER de forma escalonável. A **Figura 6-2 (a) e (b)** ilustra a diferença entre os projetos. No caso (a), a capacidade instalada é a própria capacidade da sala, o jeito antigo de fazer o projeto. No caso (b), a ideia é que a capacidade instalada aumenta em incrementos de acordo com o aumento da demanda da carga de TI.

Os novos projetos são cada vez mais baseados na opção (b) e os fornecedores estão otimizando os equipamentos utilizados no DATACENTER.

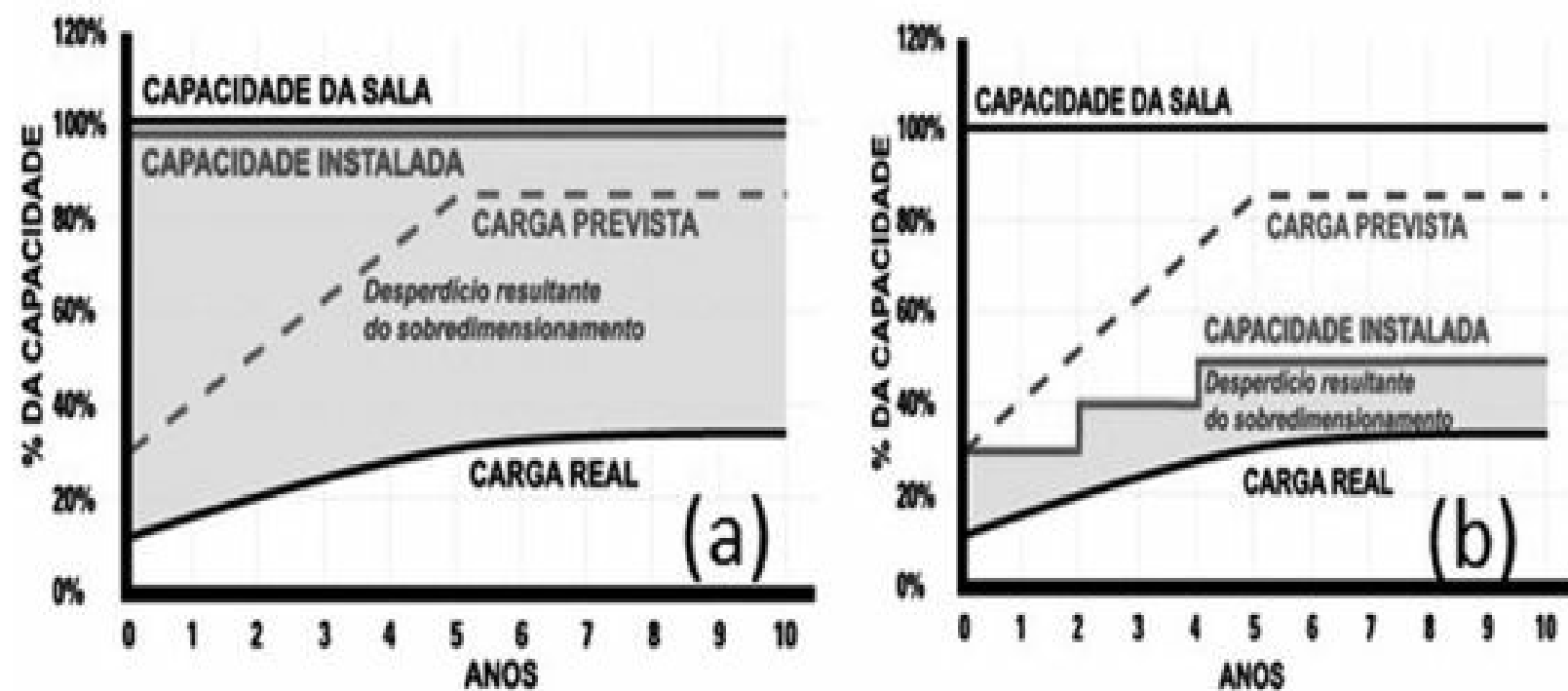


Figura 6-2 – Design da Capacidade de Energia Típico (a) e Sugerido (b) (Fonte: APC)

A APC sugere que existem cinco fatores-chave que contribuem para a ineficiência energética:

- Ineficiência dos equipamentos de energia elétrica.
- Ineficiência dos equipamentos de refrigeração.
- Consumo de energia pela iluminação.

- Superdimensionamento dos sistemas de energia e resfriamento.
- Ineficiências devido à configuração.

Segundo a APC, uma arquitetura ideal, do ponto de vista da eficiência energética, deveria utilizar os seguintes princípios:

- Equipamentos de energia e refrigeração que não são necessários no momento não devem ser energizados.
- O superdimensionamento deve ser reduzido sempre que possível para que os equipamentos possam operar na curva ideal de eficiência.
- Equipamentos de energia, refrigeração e iluminação devem aproveitar as novas tecnologias para reduzir o consumo de energia.
- Subsistemas que devem ser utilizados abaixo da capacidade nominal (devido à redundância) devem ser otimizados para aquela fração de carga e não para a eficiência com carga plena.

6.3. Equação Energética do DATACENTER

O dimensionamento das necessidades de energia de um DATACENTER requer compreensão sobre a quantidade de energia necessária para alimentar principalmente o sistema de refrigeração, os *no-breaks* e a carga de TI. Com a carga de TI prevista, é possível determinar com alguma precisão a necessidade de energia e determinar a capacidade necessária do gerador de reserva. A redundância dos sistemas utilizados é determinada pelos requisitos impostos pelas aplicações de negócio e definem a camada (TIER) do DATACENTER. Os *no-breaks* podem estar em configurações redundantes, o que reduz a eficiência energética. Estimar a energia necessária é um aspecto crucial do projeto. Se estimada para baixo, pode trazer transtornos quando da necessidade de aumento da capacidade do DATACENTER. Se estimada para cima, pode conduzir a custos iniciais elevados. Se o DATACENTER utiliza um prédio, compartilhando as instalações, este cálculo se torna mais complexo.

A **Figura 6-3** ilustra um diagrama geral de alimentação do DATACENTER. A figura ilustra o caminho que a energia (corrente pública) percorre até alimentar a carga de TI, seu objetivo.

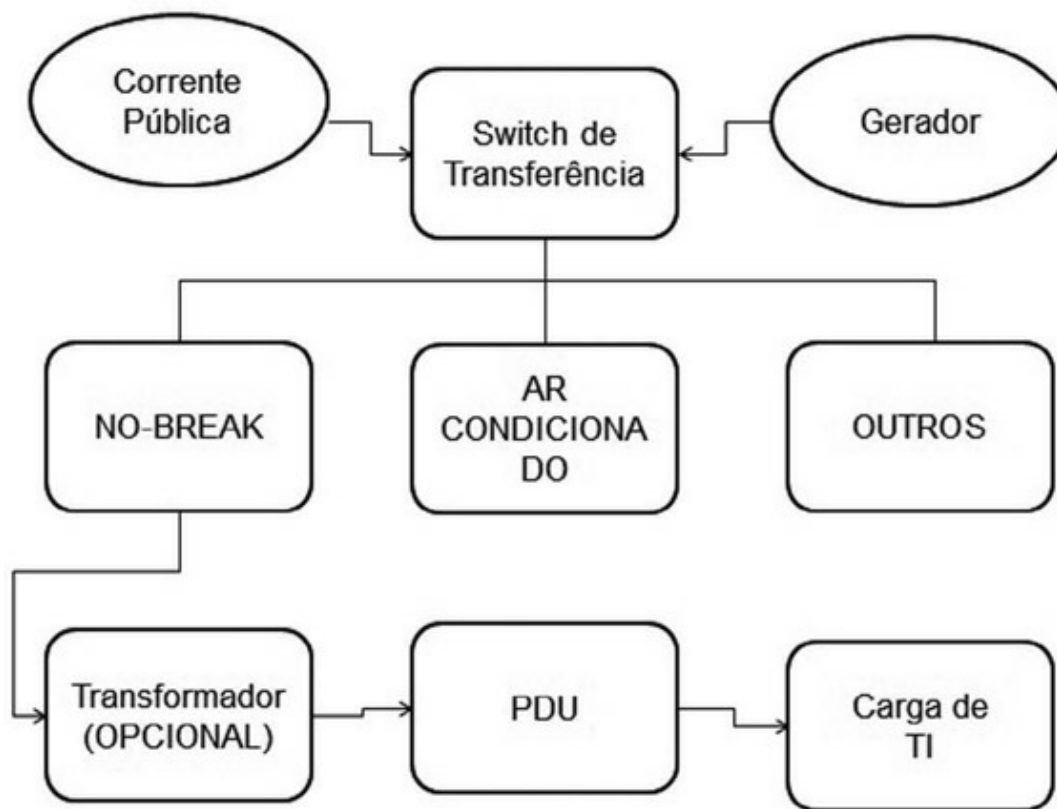


Figura 6-3 – Alimentação de um DATA CENTER (Fonte: APC)

Existem duas principais abordagens para tentar aumentar a eficiência do DATA CENTER que podem e devem ser utilizadas conjuntamente. A primeira trata de melhorar a eficiência da infraestrutura de apoio, que são os equipamentos que transformam a energia e refrigeram o ambiente. A segunda trata a eficiência no nível da carga de TI. Os principais instrumentos para atender a estas abordagens são o planejamento da capacidade instalada e a VIRTUALIZAÇÃO, descritos detalhadamente a seguir:

- **Planejar a capacidade instalada:** trata do dimensionamento correto da infraestrutura para a carga de TI esperada. Envolve conhecer melhor o consumo dos diversos equipamentos utilizados e fazer um projeto que utilize componentes escaláveis. Os projetos de DATA CENTER de alguns anos atrás consideravam a capacidade instalada muito próxima da capacidade da sala. A nova proposta para a capacidade instalada é ter equipamentos de energia e refrigeração que possam crescer (escalar) com o passar do tempo. Esta técnica possibilita uma grande redução de consumo de energia do DATA CENTER ao longo dos anos, mas muitas vezes é de difícil execução. Neste novo formato a capacidade instalada começa num patamar muito menor e cresce ao longo dos anos. É evidente que esta forma só é possível se os componentes (*no-breaks*, equipamentos de refrigeração, etc.) utilizados forem escaláveis.
- **Utilizar a VIRTUALIZAÇÃO:** a VIRTUALIZAÇÃO traz inúmeras vantagens ao ambiente de DATA CENTER. A VIRTUALIZAÇÃO de servidores, que permite o processamento em um único servidor de cargas de trabalho múltiplas, cada um com seu nível de serviço, é uma aliada na redução de calor, considerando que um servidor a plena carga ou a 15% consome praticamente a mesma energia. A possibilidade de

alocar dinamicamente estas cargas de trabalho para atender os requisitos do negócio meio que por si só já justifica a VIRTUALIZAÇÃO. Este conceito, conhecido como mobilidade da carga de trabalho, pode ser estendido para o gerenciamento térmico e de energia do DATACENTER. Em períodos de baixa utilização, parte dos servidores de um DATACENTER pode ser transicionada para um estágio de baixo consumo. Desde que a infraestrutura de energia e refrigeração possa estar integrada, esta ideia pode reduzir o consumo de energia de maneira interessante.

Outras medidas relevantes são fazer escolhas inteligentes sobre a localização de servidores, medir e monitorar a eficiência das instalações e fazer uso de melhores práticas. Também conta substituir os equipamentos legados por sistemas que oferecem alto desempenho por potência. Equipamentos antigos muitas vezes são ineficientes e em certas situações não vale a pena esperar o final do ciclo do seu uso.

A APC sugere que, para vencer os desafios impostos pela VIRTUALIZAÇÃO ao projeto de DATACENTERS, sejam adotadas as seguintes medidas:

- Para cargas de altas densidades dinâmicas, utilizar resfriamento baseado em colunas de RACKS. Em instalações convencionais de DATACENTER (corredor quente/corredor frio), fazer a retirada dos CRACs e fazer a refrigeração no próprio RACK. A solução de refrigeração no RACK varia a velocidade do ventilador de acordo com a temperatura.
- Para sistemas de energia e refrigeração superdimensionados, utilizar equipamentos de energia e refrigeração escaláveis.
- Utilizar ferramentas de gerenciamento de capacidade (*capacity management tools*) para saber se a capacidade atende a demanda no nível da fila, RACK e servidor.

Uma outra alternativa sugerida pela APC é criar uma zona de alta densidade para equipamentos do tipo *blades* com sistema exclusivo de refrigeração.

Deve ficar claro que a eficiência no consumo da carga de TI é melhorada com a VIRTUALIZAÇÃO, mas a eficiência global do DC é otimizada, adequando os equipamentos de energia e refrigeração ao consumo menor imposto pela VIRTUALIZAÇÃO.

6.4. O Green Grid

6.4.1. Introdução

O Green Grid (<http://www.thegreengrid.org/>) é um consórcio global formado por diversas companhias de TI (incluindo Intel, Dell, VMware, AMD) com o objetivo de definir e propagar melhores práticas relacionadas à eficiência no consumo de energia em DATACENTERS. O Green Grid trabalha proximamente ao consórcio EPA e ao *Alliance to Save Energy*.

Conforme visto anteriormente, o DATACENTER em operação, construído na década passada, normalmente é ineficiente do ponto de vista energético. A demanda maior de energia de servidores ultradensos como os servidores *blades* trazem um novo problema: a falta de capacidade energética

destes antigos DATACENTERS e o surgimento de pontos de alta densidade de energia dentro do DATACENTER que inviabilizam o projeto de refrigeração. Os RACKS atuais, por exemplo, podem demandar até 40KW de energia devido à densidade atual de potência, o que alguns anos atrás era impensável. Estes DATACENTERS foram projetados para máxima funcionalidade e desempenho e na época não existia tanta preocupação com o consumo de energia. Além disso, a VIRTUALIZAÇÃO permite mover as cargas de TI dentro do DATACENTER trazendo mais um componente de complexidade.

Atualmente, alguns fatores guiam os projetos dos novos DATACENTERS:

- Aumento na demanda computacional.
- Aumento na densidade dos equipamentos de TI.
- Utilização de técnicas de VIRTUALIZAÇÃO que permitem mover as cargas de TI dentro do DATACENTER.
- Aumento dos custos de energia.
- Falta de disponibilidade de energia para atender estas novas demandas.

Os custos com energia e refrigeração para servidores ao longo de três anos, por exemplo, já são maiores do que os custos de aquisição de servidores, segundo o IDC¹³.

Estes fatores tornam as métricas até então utilizadas menos importantes. O caso da métrica DCD (densidade do DATACENTER) usada por anos como a métrica principal do DATACENTER é emblemático. Esta métrica é excelente quando se quer comparar em termos absolutos o consumo de energia entre um DATACENTER e outro, mas falha quando se quer verificar a eficiência pura do DATACENTER.

$$DCD = \text{Energia consumida pelos Equipamentos} / \text{Área utilizada pelos Equipamentos}$$

Assim, a meta inicial do Green Grid foi a de definir melhores métricas para o consumo de energia do DATACENTER com o objetivo de reduzir o TCO total e ao mesmo tempo permitir a competitividade e a adequação para futuras demandas de energia.

No longo prazo o Green Grid pretende definir uma arquitetura para a definição de políticas e especificações, como também um logotipo vinculado ao *compliance* e interoperabilidade de dispositivos para o uso eficiente de energia.

6.4.2. PUE e DCiE

O Green Grid desenvolveu duas métricas táticas relacionadas e já bem utilizadas. O PUE (*Power Usage Effectiveness*) e o DCiE (*DATACENTER Efficiency*). O PUE é mais utilizado.

$$PUE = \text{Energia Total da Instalação} / \text{Energia dos Equipamentos de TI} \quad DCiE = 1 / PUE$$

A energia total da instalação é a energia medida pelo medidor que alimenta o DATACENTER.

A energia consumida pelos equipamentos de TI é a energia consumida por todos os equipamentos de TI, incluindo KVMs, monitores, estações de gerenciamento. Em muitas instalações não se consideram nem medidores separados para a carga de TI e, portanto, fica difícil medir o PUE.

A **Figura 6-4** ilustra os consumos envolvidos no cálculo do DCiE.

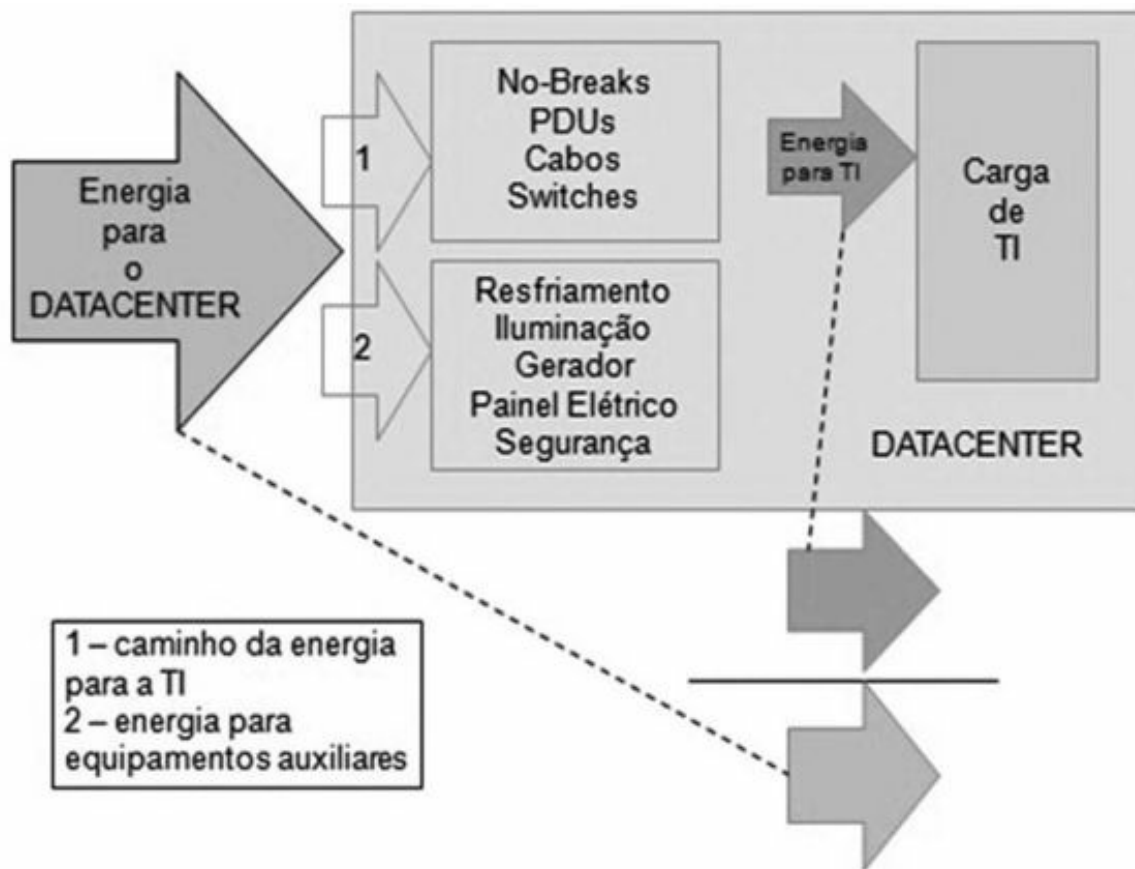


Figura 6-4 – Cálculo do DCiE em DATACENTER

Estas duas métricas permitem calcular a real eficiência do DATACENTER, comparar DATACENTERS do ponto de vista do consumo de energia e criar *benchmarks*, verificar melhoria do consumo ao longo do tempo e oportunidades para realocar energia para novos equipamentos de TI.

Um PUE de 3.0, típico de instalações antigas, indica que o DATACENTER demanda três vezes mais energia do que a necessária para alimentar os equipamentos de TI. Neste caso, se um RACK de energia demanda 10KW médios de energia para alimentar os diversos servidores, será necessário 30KW na entrada para alimentar todo o DATACENTER. Ou seja, você terá que comprar três vezes mais energia da concessionária. A eficiência é de cerca de 30%. A meta atual para novos DATACENTERS é ter um PUE o mais próximo possível de 1.

O Google fez um grande investimento na melhoria da eficiência dos seus DATACENTERS e recentemente publicou o PUE medido para dez dos seus principais DATACENTERS, todos abaixo de 1,4.

A APC, no relatório técnico 154 cujo título é “Medição da Eficiência Elétrica de DATACENTERS”, sugere a simplificação da medição do PUE. Para medir a eficiência de um DATACENTER em um ponto operacional em especial, deve-se medir a energia elétrica de entrada

total para o DATACENTER e a energia consumida pela carga total da TI.

A medição da energia total de entrada pode ser feita logo depois do switch de transferência. Este é o caso mais simples, pois o DATACENTER tem uma entrada exclusiva de energia.

O mais complicado é medir a energia consumida pela carga de TI. Se a carga for um único equipamento gigantesco, a energia elétrica da carga da TI será uma única medição na conexão elétrica do equipamento. Seriam necessárias apenas duas medições neste caso hipotético. Infelizmente, esta situação ideal nunca ocorre.

A maioria dos DATACENTERS faz parte de prédios multiuso com outras cargas, e todos os DATACENTERS são constituídos de um conjunto de equipamentos de TI – possivelmente milhares –, muitos com circuitos elétricos separados. A ideia é utilizar o conceito de carga agregada de TI simplificando a medição de consumo de diversas cargas de TI. A sugestão dada pela APC é medir a energia consumida pela carga de TI logo depois do *no-break*, mas considerar as perdas com as unidades de distribuição de energia (PDUs) introduzidas conforme ilustra a **Figura 6-5**.

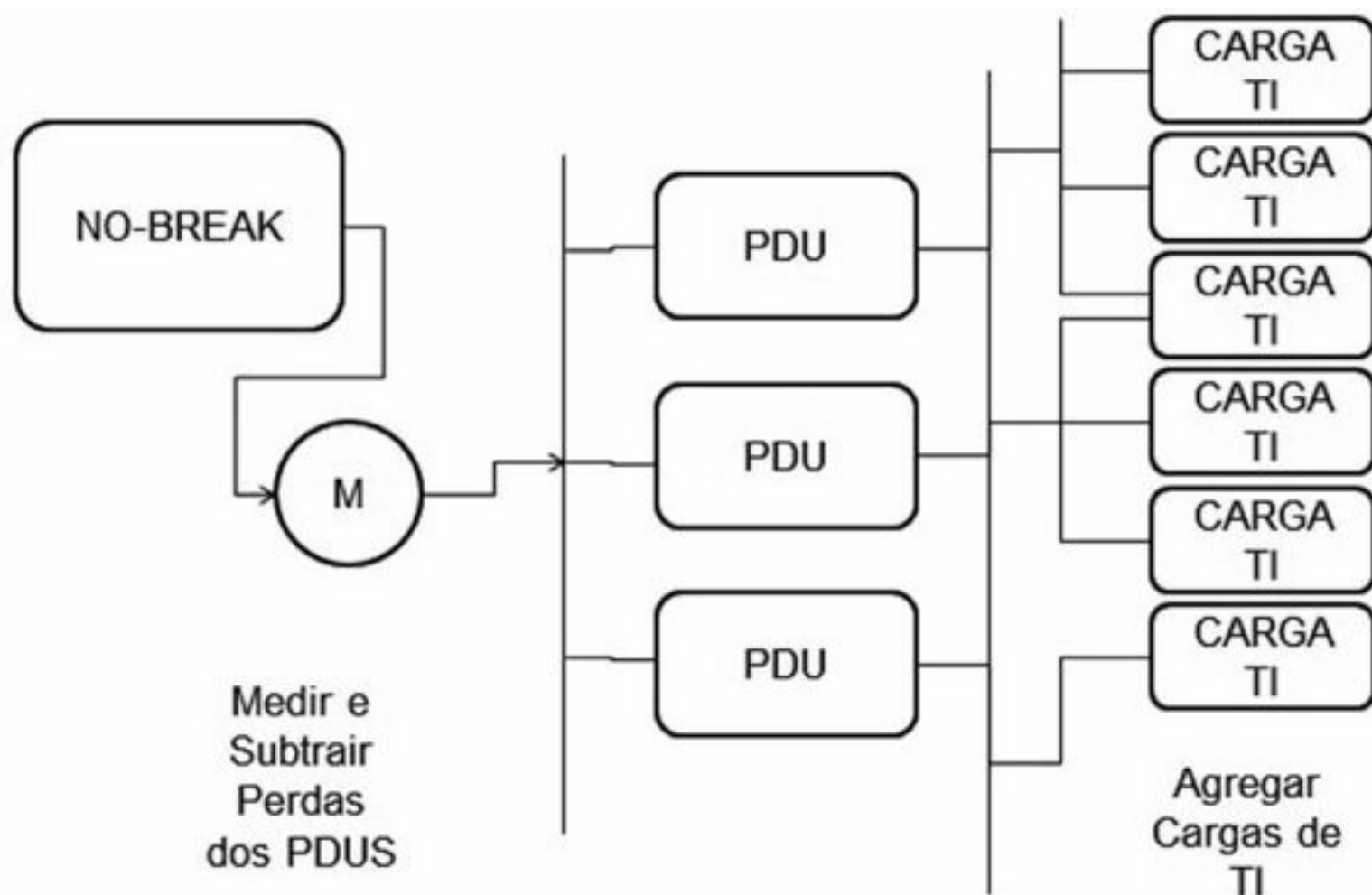


Figura 6-5 – Medição do consumo da carga de TI (Fonte: APC)

É importante ressaltar que no DATACENTER com instalações dedicadas considera-se que toda a energia consumida deve ser utilizada no cálculo do PUE. No caso de DATACENTERS compartilhados, a energia consumida por serviços auxiliares compartilhados (banheiros, escritório, etc.) deve ser excluída do cálculo.

6.4.3. PUE 2

O Green Grid continua fazendo um grande esforço para estabelecer métricas que permitam melhorar a sustentabilidade dos DATACENTERS em operação e que ajudem a aumentar a eficiência energética em novos projetos de DATACENTERS.

Diversos aspectos devem ser considerados para efeitos de melhoria da sustentabilidade de projetos de DATACENTER. Em 2007, quando o Green Grid introduziu o PUE (*Power Usage Effectiveness*), que relaciona a energia consumida pela carga de TI com a energia total consumida pelo DATACENTER, parecia que esta métrica era desconectada da realidade.

Hoje o PUE é a principal métrica relacionada à eficiência energética. Diversos artigos disponíveis na Internet contemplam o PUE, seu uso e formas de medi-lo. Grandes empresas como Microsoft e Google passaram a utilizá-lo como referência para a construção dos seus DATACENTERS. As primeiras medidas realizadas com o objetivo de medir o PUE indicavam um PUE até de 3 para boa parte dos DATACENTERS existentes. Isto de certa forma refletiu a pouca preocupação da indústria com aspectos de eficiência energética. Mais comum era a preocupação com a alta disponibilidade das instalações.

O sucesso do uso do PUE como métrica de eficiência energética por diversas organizações reforça a necessidade de surgirem novas métricas que completem a visão de sustentabilidade e diminuam os riscos e custos do ambiente.

Novos projetos utilizando novas tecnologias, maior modularidade dos componentes envolvidos no projeto e a VIRTUALIZAÇÃO permitiram que hoje existam níveis de PUE próximos à situação ideal. A própria localização do DATACENTER foi alterada na busca da melhoria da eficiência energética. Diversos DATACENTERS passaram a ser construídos em lugares frios, reduzindo assim a necessidade de refrigeração externa. Também foi feito um grande esforço na melhoria da eficiência dos equipamentos de refrigeração que respondiam pelos altos PUEs encontrados nos DATACENTERS.

Na tentativa de equacionar melhor o problema do consumo de energia por todos os componentes do DATACENTER, a Trend Point estabeleceu uma nova métrica baseada no PUE, o Micro PUE. O Micro PUE tem como objetivo avaliar as perdas de energia na refrigeração ao medir o calor que foi dissipado pelos equipamentos.

A mudança na equação do PUE se dá em discriminar a energia total da instalação em três partes: perda de eficiência na infraestrutura, consumo de energia da TI e consumo de energia da refrigeração.

Como a energia da refrigeração geralmente é maior do que as outras partes, essa relação demonstra que a redução da energia consumida pela refrigeração irá impactar mais o resultado do PUE do que outras opções. Quanto maior a energia consumida pela TI, maior será o consumo na refrigeração. Assim, o Micro PUE observa a eficiência energética em cada parte, especialmente na refrigeração.

A partir da medida de calor removido pela refrigeração, é possível determinar o consumo de energia gasto pelos equipamentos que geraram o calor no DATACENTER.

O Micro PUE demonstra a quantidade de energia usada para processar e refrigerar 1 kWh da carga da TI. É como se fosse um PUE para cada segmento de TI/refrigeração em um DATACENTER.

Com essa métrica é possível verificar a eficiência de cada unidade de resfriamento e com isso reduzir o PUE. O Micro PUE permite também observar as fontes de calor e a eficiência da respectiva resposta de refrigeração, podendo adequar a refrigeração à medida que as fontes de calor mudam. Enquanto a dinâmica de fluidos faz estimativas sobre o movimento do calor, o Micro PUE possibilita acompanhar o fluxo do calor em tempo real.

A eficiência energética em DATACENTER continua sendo estudada e ainda não foi possível chegar a uma solução definitiva. O Green Grid estabeleceu algumas convenções e princípios ao medir o PUE. Foi destacado que o PUE continua sendo a métrica mais indicada para medir a eficiência energética dos DATACENTERS. No entanto, foi apresentada uma classificação do PUE de acordo com o ponto de medição de energia. Assim, haverá uma melhor comparação entre PUE de diferentes DATACENTERS, de acordo com o grau de detalhamento obtido. Com isso, a divulgação de dados de medição de PUE deverá seguir uma nomenclatura adequada (PUE 0, PUE 1, PUE 2 e PUE 3). Nessa classificação, o uso de energia renovável ou gerada por fontes não poluentes também é levada em consideração. Nessa versão, as recomendações são limitadas à infraestrutura.

- **PUE categoria 0:** cálculo baseado na demanda, representando a carga máxima durante 12 meses de medição. A energia da TI é medida na saída do *no-break* (ou soma das saídas caso exista mais de um *no-break* instalado) durante os períodos de pico. É uma medida instantânea, não mede o impacto das mudanças na carga de TI. Essa medida só pode ser usada em DATACENTERS que não utilizam outras fontes de energia elétrica como gás natural.
- **PUE categoria 1:** cálculo baseado no consumo. A carga da TI é obtida através de medidas, durante o ano todo, na saída do *no-break*. A energia total da instalação é obtida somando-se a energia fornecida pela distribuidora e a energia obtida por outras fontes como gás natural.
- **PUE categoria 2:** cálculo baseado no consumo. A carga da TI é obtida através de medidas, durante o ano todo, na saída da unidade de distribuição. A energia total da instalação é obtida da mesma forma da categoria 1. Essa medida está mais próxima do consumo da TI, pois é realizada após as perdas ocorridas durante o processo de conversão existente.
- **PUE categoria 3:** cálculo baseado no consumo. A carga da TI é obtida através de medidas, durante o ano todo, no ponto de conexão dos dispositivos de TI à rede elétrica. A energia total da instalação é obtida da mesma forma da categoria 1. Essa medida oferece a maior precisão quanto ao consumo da carga da TI.

A **Tabela 6-1** exibe o resumo da classificação do PUE. No caso de outras fontes energéticas também alimentarem o DATACENTER, isso deverá ser levado em consideração no cálculo do PUE.

Tabela 6-1 – Classificação do PUE

	PUE0	PUE1	PUE2	PUE3
Local de medição da energia da TI	Saída do <i>no-break</i>	Saída do <i>no-break</i>	Saída da unidade de distribuição	Entrada do servidor
Energia da TI	Picos de demanda energética	Energia anual da TI	Energia anual da TI	Energia anual da TI
Energia total da instalação	Energia total dos picos de demanda	Energia anual do DATACENTER	Energia anual do DATACENTER	Energia anual do DATACENTER

FONTE: Green Grid (2010)

Também foram definidos pesos para as diversas fontes de energia. Os pesos mostrados na **Tabela 6-2**, definidos pelo Green Grid, foram obtidos com base em dados da EPA. Todas as fontes de energia devem ser convertidas para a mesma unidade antes de aplicar o peso da energia e somar com as outras fontes.

Tabela 6-2 – Pesos para fontes de energia

Tipo de Energia	Peso
Eletricidade	1.0
Gás natural	0.31
Óleo combustível	0.30
Outros óleos	0.30
District chilled water	0.31
District hot water	0.40
District steam	0.43
Condenser water	0.03

As energias renováveis (RE) são consideradas fora do limite de cálculo do PUE e devem utilizar o mesmo peso da eletricidade. Essa distinção ocorre em virtude da necessidade de que o PUE seja independente da fonte de energia utilizada. Caso o DATACENTER use fontes de energia limpas, com o gás natural (*Combined Heat and Power Plants* – CHP), isso deverá reduzir o PUE.

As medições que caracterizam o PUE 2 devem considerar o ponto de medição, a frequência de medição e a duração da medida.

6.4.4. Novos Indicadores CUE e WUE

Recentemente o Green Grid introduziu mais duas métricas relacionadas à sustentabilidade do DATACENTER. São métricas que no futuro serão cada vez mais utilizadas.

A primeira é o CUE (*Carbon Usage Effectiveness*), que endereça a emissão de carbono de DATACENTERS. A segunda é o WUE (*Water Usage Effectiveness*), que endereça o uso da água em DATACENTERS.

Juntas, as três métricas, PUE, CUE e WUE, devem ter cada vez mais importância e serem consideradas na localização, no projeto e na operação dos DATACENTERS.

Estas métricas permitem melhorar a sustentabilidade dos projetos, permitem fazer comparação entre projetos similares, reforçam a importância de utilizar fontes de energia renováveis, etc.

CUE (Carbon Usage Effectiveness)

Por endereçar a questão crítica da emissão de carbono, seu uso junto com o PUE permite a realização de projetos sustentáveis.

CUE = Emissão de CO₂ causada pela energia total do DATACENTER / Energia do Equipamento de TI

Em (kgCO₂eq) por kilowatt-hora (kWh)

Uma alternativa para definir o CUE é multiplicar o fator de emissão de carbono (CEF) pelo PUE anual do DATACENTER.

CUE = CEF x PUE

O CEF é o fator de emissão de carbono e é dado em kgCO₂eq/kWh. O CEF é baseado em dados publicados pelo governo na região de operação para o ano inteiro. PUE é o PUE anual.

A energia dos equipamentos de TI inclui o consumo de todos os equipamentos de TI.

A energia total do DATACENTER inclui todo o consumo da TI e dos equipamentos que suportam a TI.

A emissão total de CO₂ inclui as emissões do local e das fontes de energia utilizadas.

Importante ressaltar que as medidas de emissão de CO₂ são médias e são anuais e que emissões de outros gases como CH₄ precisam ser convertidas para emissão de CO₂.

WUE (*Water Usage Effectiveness*)

A métrica utilizada para avaliar o uso da água no DATACENTER em alto nível é:

$$WUE = \text{Uso anual da água} / \text{Energia do Equipamento de TI}$$

WUE é dado litros/kilowatt-hora (L/kWh).

O consumo de água no DATACENTER é uma questão complexa. Em certas situações reduzir o consumo de água pode aumentar o consumo de energia.

Se o DATACENTER faz opção por utilizar um *chiller* de expansão direta (DX) que não necessita de água e não utiliza um *chiller* que usa a evaporação de água como mecanismo de rejeição do calor, pode-se num primeiro momento achar que se está otimizando o WUE, mas na verdade pode-se estar aumentando a necessidade de energia para os equipamentos que suportam a TI.

A água considerada na fórmula é a água usada na operação do DATACENTER, incluindo umidificação, a água usada para refrigerar o DATACENTER e a água utilizada na produção de energia.

O Green Grid também define uma métrica que inclui a água usada *on-site* e a água usada *off-site* na produção da energia usada *on-site*. Tipicamente, isto adiciona a água usada na geração de energia na fonte para a água usada no site.

$$WUE_{source} = EWIF \times PUE \times WUE$$

Onde $EWIF = \text{Energy Water Intensity Factor}$ é necessário para calcular WUE_{source} . $EWIF$ é o volume de água usado para gerar energia.

O consumo de água na geração da energia de um DATACENTER na fonte deve ser considerada para atividades como seleção de site e projeto do sistema.

As três métricas PUE, CUE e WUE usam como denominador o consumo de energia da carga de TI. O valor ideal para o PUE é 1, já para o CUE e o WUE é 0. CUE de zero implica em não ter emissão de carbono associada à operação do DATACENTER. WUE de zero implica em não haver consumo de água quando da operação do DATACENTER. Estas métricas não consideram o ciclo completo do DATACENTER e, por exemplo, não cobrem as emissões de carbono ou a água gasta quando da fabricação dos equipamentos a serem utilizados pelo DATACENTER.

6.4.5. Modelo de Maturidade do DATACENTER (DCMM)

Melhorar a eficiência e a sustentabilidade de DATACENTERS é o que se busca. O DATACENTER Maturity Model (DCMM) possibilita que donos de DATACENTERS e operadores possam ter um *benchmark* de sua performance atual, possibilita determinar seu nível de maturidade e possibilita identificar passos a serem dados para melhorar a sustentabilidade e a eficiência energética.

A força tarefa criada pelos integrantes do Green Grid para criar o modelo DCMM definiu seis níveis que variam de 0 a 5. Vale ressaltar que este modelo ainda é imaturo e deverá ser aprimorado com o passar do tempo.

- 0 – mínimo.
- 1 – melhor prática parcial.
- 2 – melhor prática.
- 3 e 4 – entre melhores práticas correntes e visionário 5 anos à frente.
- 5 – visionário – 5 anos à frente.

O Green Grid espera que organizações que estejam no nível 2 em 2011 estejam no nível 5 em 2016.

Os autores do modelo alertam que as melhorias devem ser feitas na TI também.

As áreas envolvidas no modelo DCMM são:

■ Instalações

- ☐ Energia
- ☐ Refrigeração
- ☐ Gerenciamento
- ☐ Outros

■ TI

- ☐ Processamento
- ☐ Armazenamento
- ☐ Redes
- ☐ Outros

A **Figura 6-6** ilustra a utilização do modelo para o item gerenciamento. Para cada uma das variáveis de gerenciamento verifica-se o estado atual comparando com as melhores práticas do setor. Pode-se então traçar um plano de melhorias para o item gerenciamento. Pode-se fazer o mesmo com os outros itens. Maiores detalhes do modelo podem ser encontrados no próprio site do Green Grid.

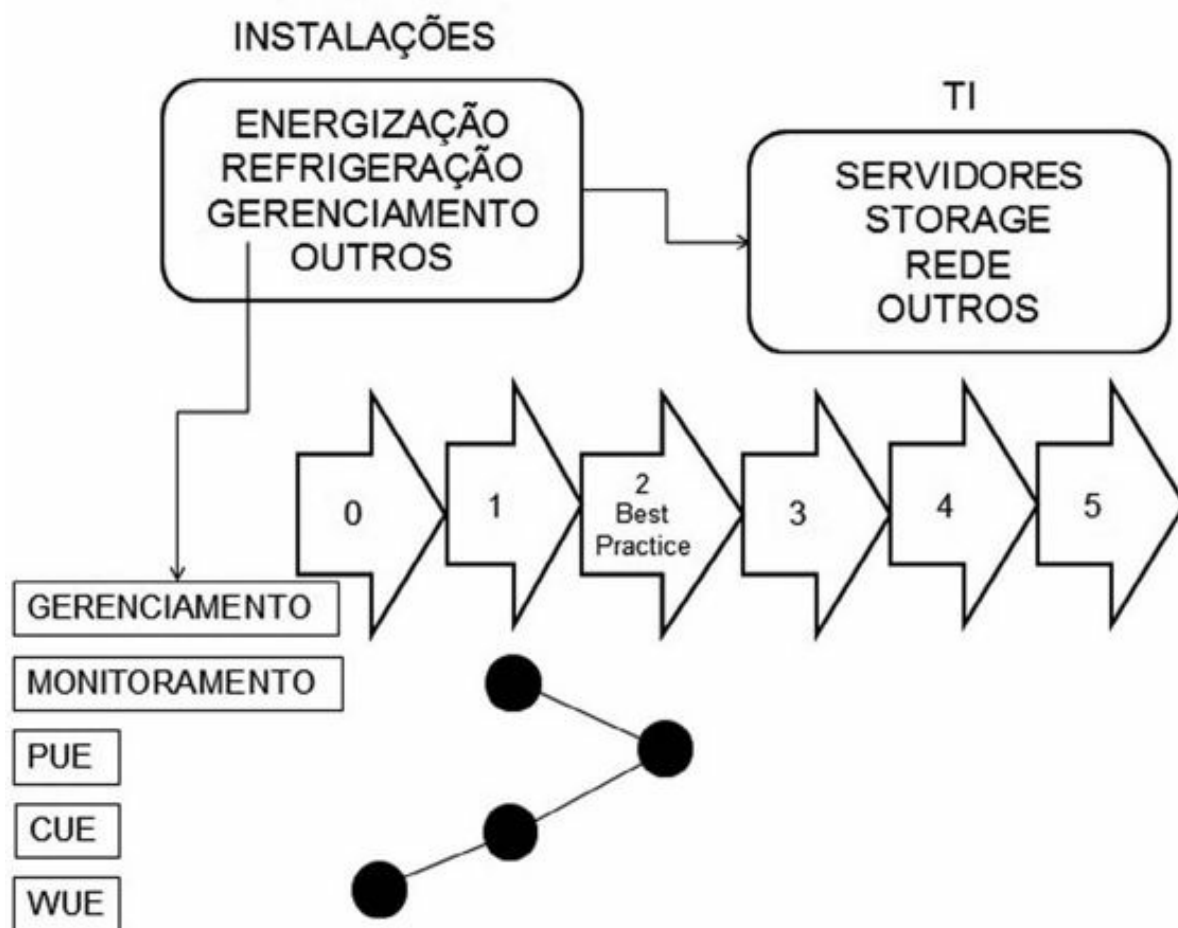


Figura 6-6 – A ideia do DCMM

6.5. Conceito na Prática: Open Compute Project

O Open Compute Project, já mencionado no capítulo 5, é um conjunto de tecnologias que se propõe a reduzir o consumo de energia e o custo de DATACENTERS, aumentando a confiabilidade e a escolha no mercado e ainda simplificando a operação e manutenção. O projeto abre as especificações elétricas e mecânicas para os principais componentes do DATACENTER.

O DATACENTER do Facebook em Oregon foi o local da primeira implementação destas tecnologias. Os principais componentes deste projeto considerado aberto são as especificações elétricas e mecânicas.

As instalações utilizam sistemas elétricos baseados em *no-break* de 48VDC integrado com uma fonte de 277VAC para os servidores. Os sistemas mecânicos utilizam 100% *airside economizers* com um sistema de refrigeração do tipo evaporativo. As fontes dos servidores também fazem parte das especificações e possuem 95% de eficiência e 48VDC de energia de backup.

Esta especificação foi disponibilizada com o nome de *Open Web Foundation Specification Agreement* (OWFa 1.0) e está disponível na web.

6.6. Conceito na Prática: Google DATACENTERS

O Google, em função do seu negócio de busca, tem feito investimentos enormes na construção

de DATACENTERS. A busca da eficiência energética virou um padrão e outras empresas seguiram o mesmo caminho. Novos DATACENTERS do Google já apresentam PUE da ordem de 1.1.

Recentemente, a empresa recomendou uma série de melhores práticas que aqui são resumidas.

- **Medir o PUE:** o Google aconselha medir o PUE durante todo o ano.
- **Gerenciar o fluxo de ar:** o Google aconselha gerenciar o fluxo de ar eliminando *hot spots*. Modelamento via CFD pode ajudar a caracterizar e otimizar o fluxo de ar.
- **Ajustar a temperatura:** o Google orienta a utilização de temperaturas maiores dentro do DATACENTER para facilitar o uso de *air-side economizers*.
- **Usar refrigeração livre:** a ideia é remover calor sem utilizar o *chiller*. Isto pode ser feito utilizando uma temperatura ambiente baixa, evaporação da água ou usando um grande reservatório térmico. A utilização de *economizers* a água ou a ar é uma medida que melhora a eficiência energética.
- **Otimizar a distribuição de energia:** minimizar as perdas de distribuição de energia eliminando passos de conversão quando possível. PDUs e transformadores eficientes fazem parte desta medida. O *no-break* deve ser muito eficiente. Também o Google recomenda utilizar equipamentos com fontes de energia com alta eficiência, incluindo servidores.

6.7. Conceito na Prática: HP POD 240a

Diversos negócios querem implementar DATACENTERS de forma mais rápida, com custos menores e mais eficientes. Segundo a HP, DATACENTERS modulares são mais interessantes do que DATACENTERS monolíticos (*Brick and Mortar*). A [Figura 6-7](#) ilustra o novo DATACENTER da HP, o HP POD 240a.



O HP 240a utiliza a *Adaptive Cooling Technology* (ACT), que otimiza a eficiência e reduz a emissão de carbono com um sistema de refrigeração que se adapta automaticamente.

No HP 240a a solução de TI pode ser testada na fábrica.

O HP 240a pode ser implementado em 12 semanas, contra 24 meses de DATACENTERS tradicionais.

Segundo a HP, o HP240a é 25% do custo de um DATACENTER convencional para uma mesma potência (2.3MW com densidade de energia nos RACKS de 44KW) e é 95% mais eficiente. Pode utilizar até 44 RACKS de 50U e integra energia, refrigeração, monitoramento e supressão de incêndio e fogo. O PUE do HP240a varia de 1.03 (refrigeração livre), 1.15 (refrigeração do tipo *direct assist*) até 1.3 (*full DX direct expansion*), em função do tipo de refrigeração que está sendo utilizado no momento. Segundo a HP, o HP 204a é TIER 3.

6.8. Referências Bibliográficas

- Best Practices for Unlocking your Hidden DATACENTER, Dell Power Solution, february 2008.
- Calcular requisitos de potência totais em centros de dados, AT 3 revisão 1, APC.
- Cálculo dos requisitos de refrigeração total para central de dados, APC Relatório Interno 25, APC.
- Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric. The Green Grid, 2010.
- DATACENTER Efficiency in the Scalable Enterprise, Dell Power Solutions, 2007.
- Deploying High-Density Zones in a Low-Density DATACENTER. APC White Paper #134.
- Evitando despesas por obter um DATACENTER e Sala de Infraestrutura de Rede maior que o necessário. APC White Paper #37.
- Green Grid Metrics: Describing DATACENTER Power Efficiency. The Green Grid, 2007.
- Medição da Eficiência Elétrica de DATACENTERS, APC White paper #154.
- O DATACENTER verde, IBM, 2007.
- Recommendations for Measuring and reporting Overall Data Centre Efficiency, Version 2 – Measuring PUE for Data Centers, 17 may 2011.
- The Green Grid Opportunity: decreasing DATACENTER and other IT energy usage patterns, The Green Grid, February 2007.
- Uma arquitetura aperfeiçoada para DATACENTERS de alta densidade e alta eficiência. APC White Paper #126.
- Veras, Manoel. DATACENTER: Componente Central da Infraestrutura de TI, Brasport, 2009.
- Virtualization: Optimized Power and cooling to Maximize Benefits. APC White Paper #118.
- Water Usage Effectiveness (WUE): A Green Grid Data Center Sustainability Metric, Green Grid, 2011.

7. DATACENTER: Arquitetura

7.1. Introdução

A arquitetura de TI, definida aqui como a arquitetura dos aplicativos e a arquitetura da infraestrutura, é parte importante do projeto dos DATACENTERS que vão compor a nuvem. A Microsoft, por exemplo, partiu para construir seus DATACENTERS definindo suas arquiteturas e *form factors* de forma padronizada. A arquitetura da infraestrutura, por sua vez, tem como componente principal o DATACENTER, que também é constituído por diversas arquiteturas.

A melhor forma de entender a estrutura hierárquica do DATACENTER é utilizar o modelo da norma TIA-942 sugerido na [Figura 7-1](#).

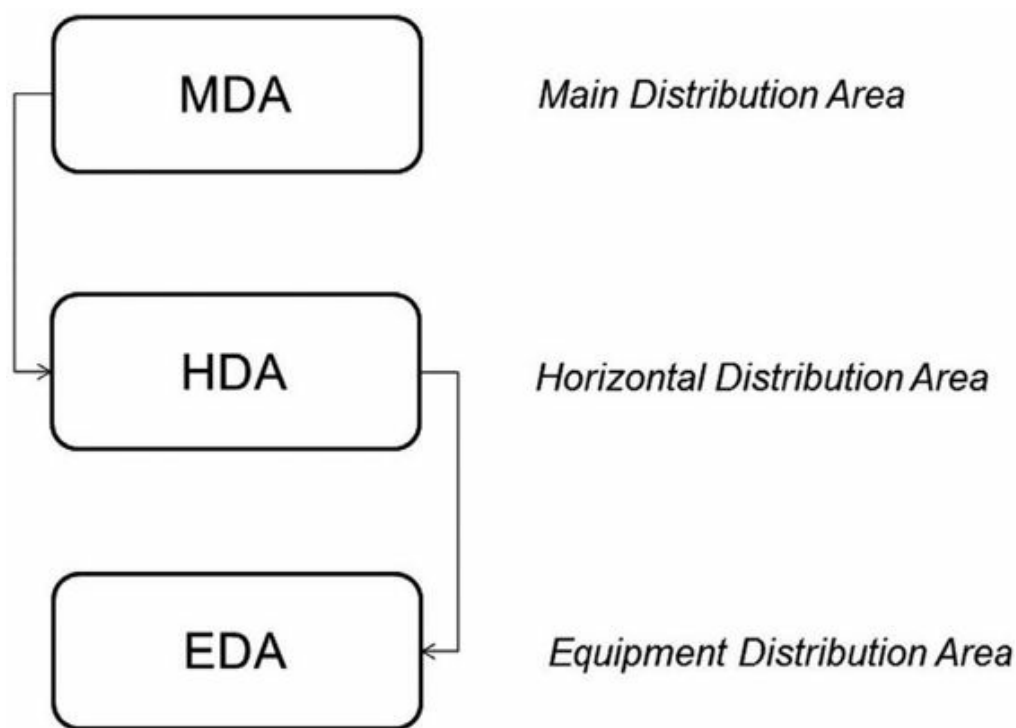


Figura 7-1 – Arquitetura e TIA-942

7.2. Arquitetura do DATACENTER

Normalmente o DATACENTER é construído com base em uma arquitetura tecnológica hierárquica que segue o modelo sugerido pela Cisco. O modelo é o utilizado em projetos de redes, só alterando a camada de distribuição pela camada de agregação. Esta forma de construir a arquitetura da infraestrutura do DATACENTER sugerida pela Cisco está de acordo com a arquitetura proposta

pela norma TIA-942.

A **Figura 7-2** ilustra a arquitetura em camadas do DATACENTER.

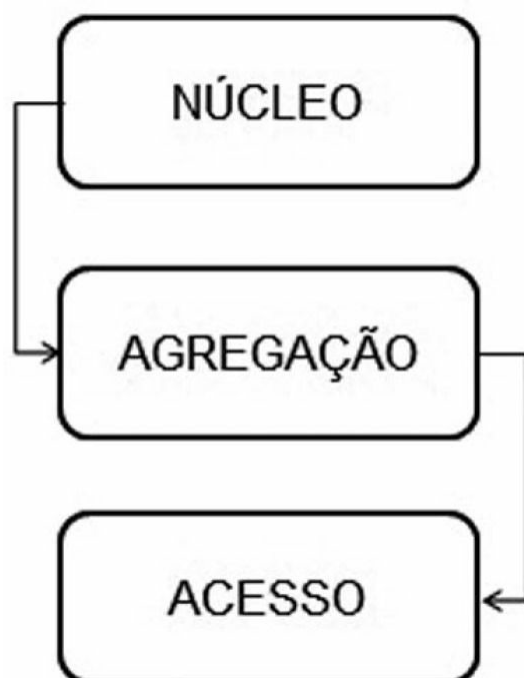


Figura 7-2 – Camadas no DATACENTER

A Cisco afirma que esta arquitetura baseada em camadas tem sido testada e validada por vários anos em grandes DATACENTERS espalhados pelo mundo.

As camadas de rede do modelo hierárquico são:

- **Núcleo:** formada por switches de alta velocidade para os fluxos que entram e saem do DATACENTER.
- **Agregação:** formada por switches com função de integração dos principais serviços de rede incluindo balanceamento de carga, detecção de intrusão, firewalls, SSL *offload*, etc.
- **Acesso:** formada por switches onde os servidores são conectados à rede e onde as políticas de rede (ACLs, QoS, VLANs) são implementadas.

A camada de acesso pode ser implementada com grandes switches modulares localizados no fim de cada coluna localizada no DATACENTER (Modelo *End-of-Row* – EoR) ou com switches com configuração fixa no topo dos RACKS (Modelo *Top-of-Rack* – ToR) que propiciam conectividade para um ou alguns RACKS adjacentes e possuem link para a camada de agregação. É possível utilizar cabos de cobre categorias 6 e 6a para redes em 10 Gbps. Para redes em 40 Gbps e 100 Gbps é necessário utilizar cabos de fibra óptica categorias OM3 e OM-4 (ambos multimodo) e OS2 (mono modo).

De acordo com estudo publicado nos Estados Unidos, a proporção de uso de fibra e cobre nos DATACENTERS nos EUA era de 52% (Fibra) e 58% (Cobre) em 2008.

Segundo o mesmo estudo, os principais critérios para seleção eram o custo e o desempenho.

O avanço da tecnologia Ethernet tem possibilitado a unificação das redes dentro do DATACENTER. Uma tendência é a de que toda a comunicação interna seja baseada em dispositivos que se conectam via Ethernet. A previsão de aumento de velocidade deste padrão de 1GbE para 10GbE, 40GbE e 100GbE nos próximos dez anos reforça e sua utilização para conectar servidores.¹⁴

Os principais recursos disponibilizados por uma arquitetura do DATACENTER físico são:

- Recursos de processamento e memória, incluindo servidores e *clusters*.
- Recursos de armazenamento.
- Recursos de conectividade.

Dois efeitos recentes alteram o perfil de arquitetura e a forma de disponibilizar recursos até então utilizados no DATACENTER.

- A introdução da arquitetura baseada em *blades* produz mudanças profundas no DATACENTER, alterando o perfil de utilização do espaço físico, o gerenciamento, a comunicação e a refrigeração.
- A VIRTUALIZAÇÃO invalida ou pelo menos altera a arquitetura convencional mostrada anteriormente, pois altera a relação de uma imagem de sistema operacional para um servidor (1:1) para várias imagens para um servidor (n:1).

7.3. Arquitetura Virtual do DATACENTER

A VIRTUALIZAÇÃO possibilita otimizar o uso da infraestrutura de TI incluindo servidores, *storage* e os dispositivos de rede. Ela agrega os vários recursos e apresenta um simples e uniforme conjunto de elementos em um ambiente virtual. O DATACENTER virtual pode então ser provisionado para o negócio com o conceito de infraestrutura compartilhada virtualmente.

Os principais recursos da arquitetura do DATACENTER virtual são:

- Recursos de processamento e memória incluindo servidores, *clusters* e RPs (*Resource Pools* ou pool de recursos).
- Recursos de armazenamento e *datastores*.
- Recursos de conectividade.

Os servidores virtuais representam os recursos virtuais de processamento e memória de um servidor físico rodando o software de VIRTUALIZAÇÃO. No servidor estão alojadas as máquinas virtuais (VMs).

Os *clusters* virtuais são conjunto de servidores que possibilitam agregar dinamicamente recursos de processamento e memória. Estes *clusters* atuam de forma coordenada e permitem obter funcionalidades que não são obtidas com um servidor isolado.

Os recursos de processamento e memória incluídos em servidores e *clusters* podem ser

particionados em uma hierarquia de pools de recursos.

No *storage* aplicações, dados e informações de configuração ficam armazenados. Os *datastores* são representações virtuais de combinações de recursos físicos de *storage*.

As máquinas virtuais (VMs) são associadas a um servidor virtual particular, um *cluster*, um pool de recursos e um *datastore*. Provisionar VMs é mais simples do que provisionar servidores físicos. Os recursos são provisionados para as VMs com base nas políticas definidas pelo administrador de sistemas, que assim permite obter uma grande flexibilidade no novo ambiente. O pool de recursos pode ser reservado para uma VM específica, por exemplo. O pool de recursos permite alocar recursos específicos para as aplicações e funcionam dentro de um *cluster* das máquinas virtuais.

O DATACENTER virtualizado é um conjunto de tecnologias (*computing pods*) que funcionam como grandes conjuntos de recursos, conforme ilustra a **Figura 7-3 (a)**. Cada pool de recursos é representado por um conjunto de BLADES, *storage* e rede. Para cada um destes conjuntos existe um sistema de energia e refrigeração projetado para otimizar o uso dos recursos. Para cada conjunto de tecnologias existe uma plataforma simplificada de gerenciamento e provisionamento. Os módulos tipo CONTAINERS, ou módulos de DATACENTERS convencionais, podem ser os elementos físicos onde os PODs estão instalados, conforme ilustra a **Figura 7-3 (b)**. Um conjunto de CONTAINERS forma o DATACENTER, conforme ilustra a **Figura 7-3 (c)**.

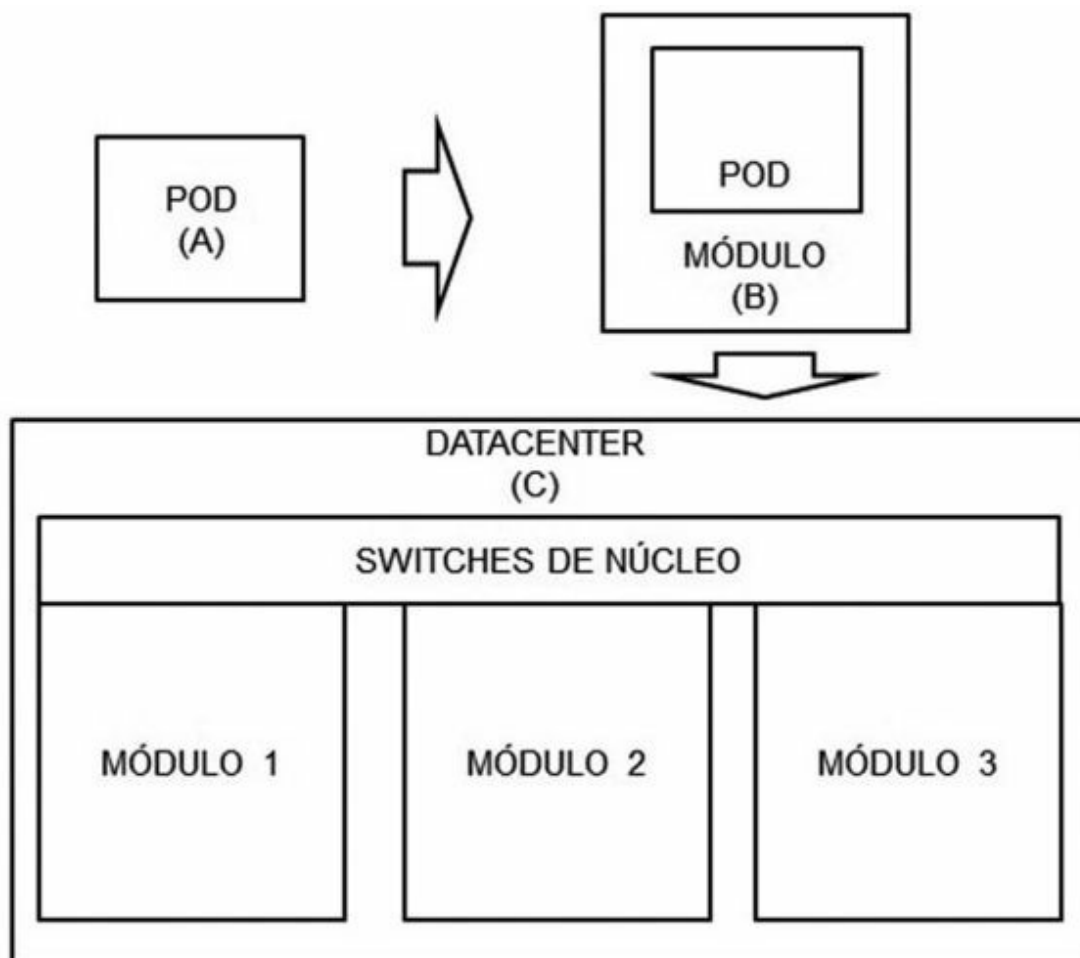


Figura 7-3 – PODs, Módulos e DATACENTER

Na arquitetura convencional os servidores rodam sistemas operacionais exclusivos que utilizam dispositivos de I/O físicos ligados ao servidor. A VIRTUALIZAÇÃO modifica esta arquitetura, na medida em que invalida a ideia central de um servidor físico rodando um único sistema operacional.

A VIRTUALIZAÇÃO também invalida a ideia de que a relação entre o sistema operacional e a rede é estática. Como a VIRTUALIZAÇÃO abstrai o hardware do software, ela permite que a imagem do SO, agora uma máquina virtual, possa ser movimentada entre servidores ou mesmo entre DATACENTERS. Do ponto de vista da rede, significa que os serviços implementados na camada de agregação precisam ser modificados para suportar a mobilidade das máquinas virtuais. Mesmo na camada de acesso, a mobilidade das máquinas virtuais pode trazer algumas novas necessidades que vão ao encontro das melhores práticas de redes estáticas.

7.4. VIRTUALIZAÇÃO

7.4.1. Conceito

A VIRTUALIZAÇÃO pode ser conceituada de duas principais formas:

É o particionamento de um servidor físico em vários servidores lógicos. Na **Figura 7-4** cinco servidores são substituídos por um servidor virtual e, portanto, a taxa de consolidação é de 5:1.

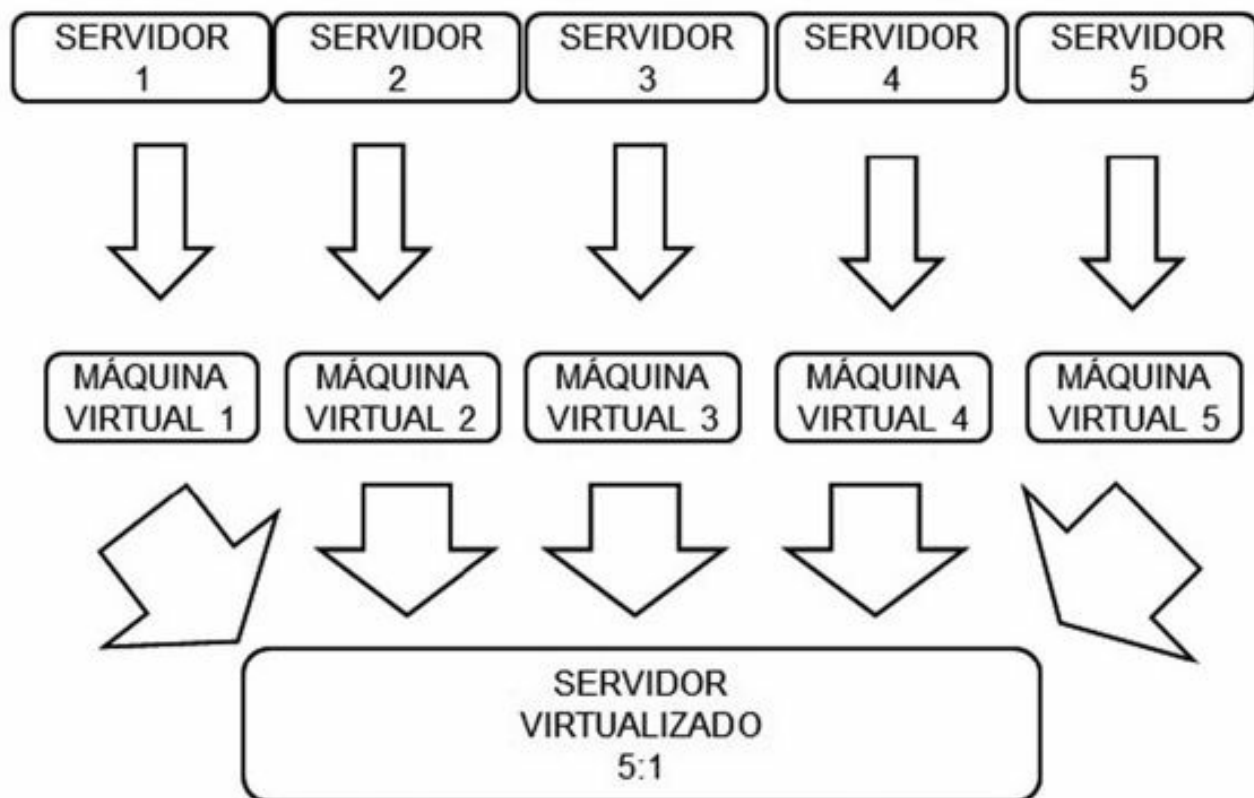


Figura 7-4 – O que é a VIRTUALIZAÇÃO

- É uma camada de abstração entre o hardware e o software que protege o acesso direto do software aos recursos físicos do hardware. A VIRTUALIZAÇÃO permite que a camada

de software (aplicações e sistema operacional) seja isolada da camada de hardware. Normalmente a VIRTUALIZAÇÃO é implementada por um software. A **Figura 7-5** ilustra o conceito.

A VIRTUALIZAÇÃO simplifica o gerenciamento, permite flexibilizar e ampliar o poder de processamento. Funcionalidades contidas nos softwares de VIRTUALIZAÇÃO também permitem melhorar a disponibilidade e a recuperação de desastres de ambientes de TI de uma maneira mais simples e com menor custo quando comparado a formas tradicionais.

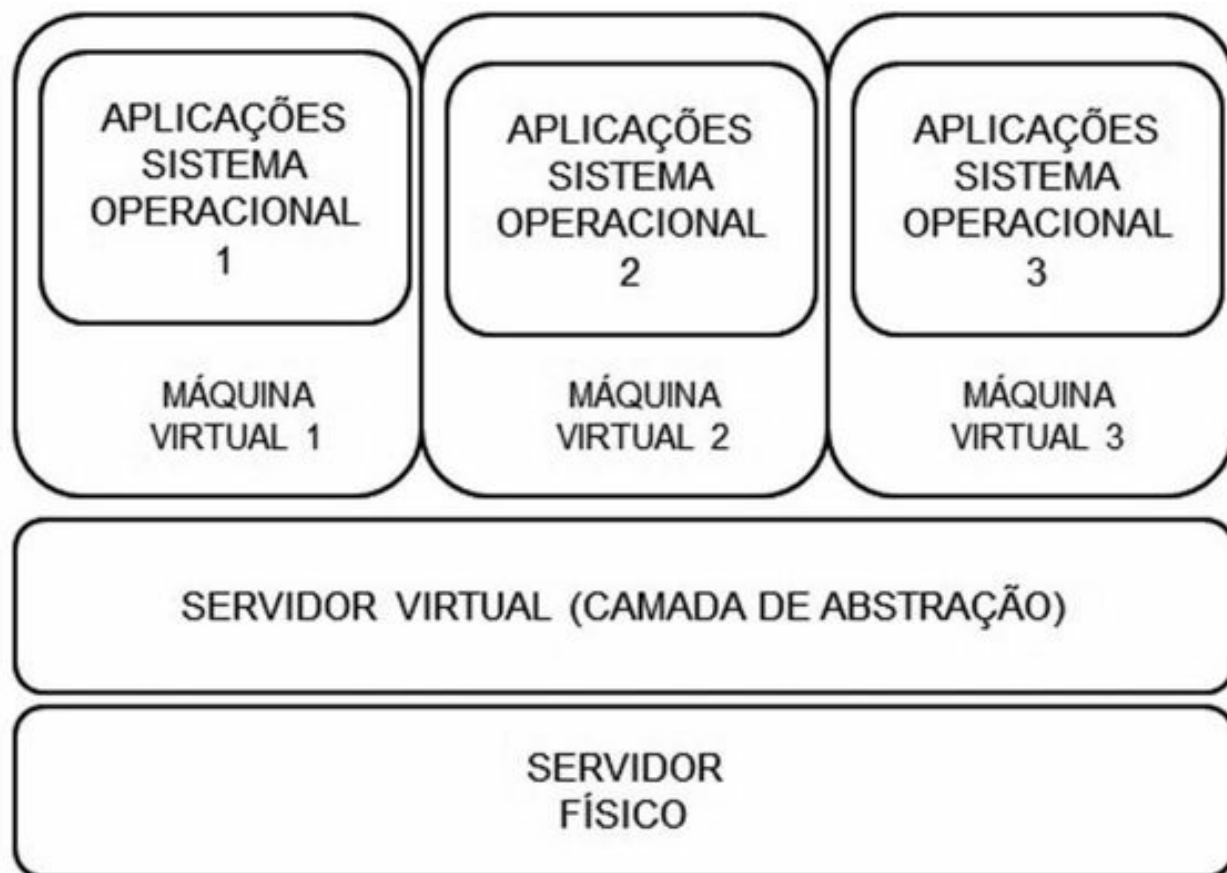


Figura 7-5 – Arquitetura da VIRTUALIZAÇÃO

A camada de VIRTUALIZAÇÃO entrega para o sistema operacional convidado um conjunto de instruções de máquinas equivalente ao processador físico. A camada de VIRTUALIZAÇÃO de servidores mais conhecida é o **HYPERVISOR** ou Monitor de Máquina Virtual (*Virtual Machine Monitor – VMM*). O servidor físico virtualizado pode então rodar vários servidores virtuais chamados de máquinas virtuais (*virtual machines – VMs*).

Com a VIRTUALIZAÇÃO, cada VM utiliza um sistema operacional e suas respectivas aplicações, e diversas VMs podem coexistir no mesmo servidor físico.

A **Figura 7-6** ilustra a VIRTUALIZAÇÃO com **HYPERVISOR**.

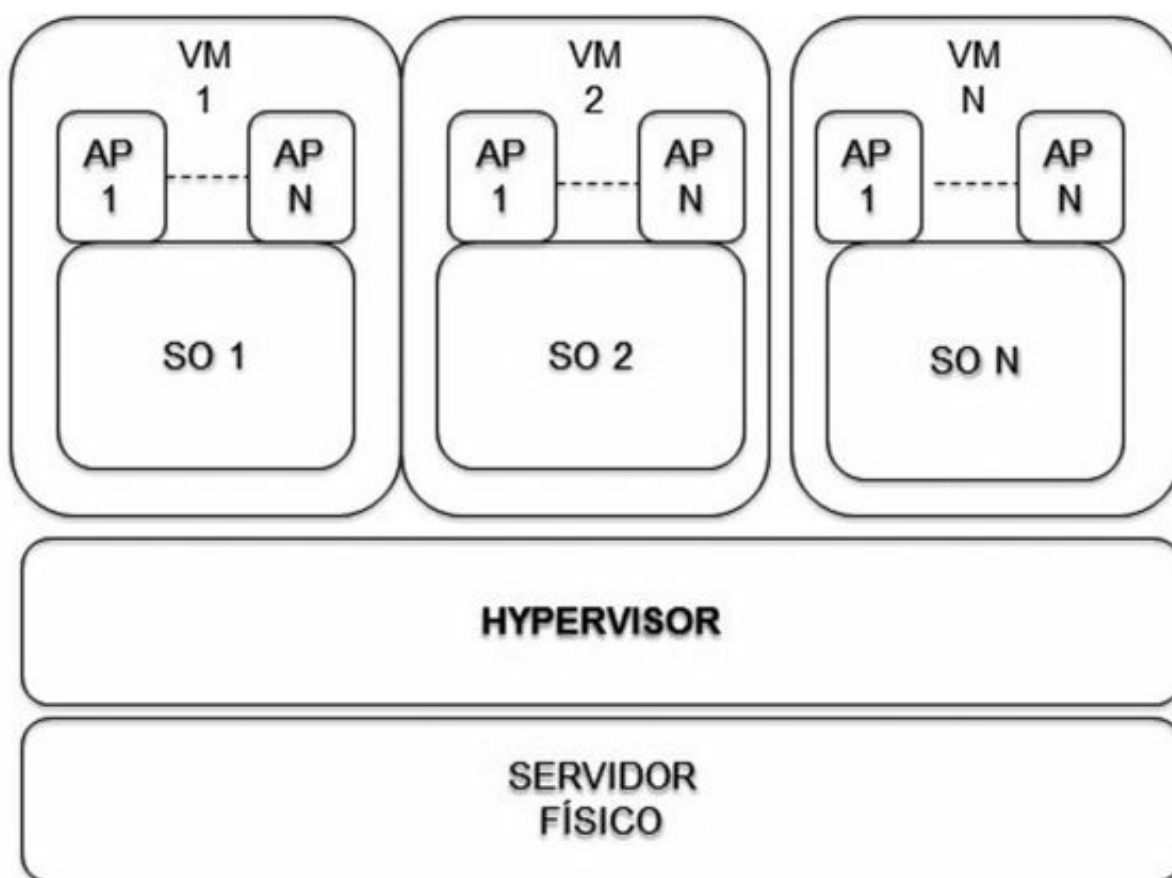


Figura 7-6 – VIRTUALIZAÇÃO com HYPERVISOR

7.4.2. Efeitos

O software de VIRTUALIZAÇÃO exige uma arquitetura e uma infraestrutura de apoio que podem ser muito simples em pequenas instalações, mas também podem ser muito complexas em grandes instalações.

Os principais softwares do mercado, como o VMware VSphere e o Microsoft WS08 Hyper-V, demandam uma infraestrutura que deve permitir obter os níveis de serviço adequados para as aplicações em termos de desempenho, disponibilidade e segurança. Como regra geral, o software de VIRTUALIZAÇÃO deve otimizar a demanda de processamento, de armazenamento e de I/O, distribuindo a carga de trabalho visando otimizar o uso dos recursos.

A infraestrutura de TI existente deve ser avaliada em novos projetos de VIRTUALIZAÇÃO. Alguns componentes da infraestrutura, como servidores e unidades de armazenamento já utilizados pela organização, podem não ser aproveitados em um projeto de VIRTUALIZAÇÃO devido a aspectos de interoperabilidade ou mesmo homologação da solução. Também quando da migração para o novo ambiente, deve-se considerar a existência de ferramentas que automatizam a migração das máquinas físicas (P) para máquinas virtuais (V), a chamada operação P2V. Estas ferramentas são disponibilizadas pelos principais fabricantes de software. Também a definição de métricas de desempenho é uma questão relevante para monitoração da qualidade da infraestrutura de TI antes e depois da VIRTUALIZAÇÃO.

Um projeto de VIRTUALIZAÇÃO traz implicações para os outros componentes da TI e para a própria arquitetura utilizada no DATACENTER. A VIRTUALIZAÇÃO altera alguns pressupostos comuns em um projeto convencional e introduz modificações profundas na maneira de pensar o

projeto do DATACENTER.

Também a VIRTUALIZAÇÃO traz impactos para o projeto da engenharia do DATACENTER, pois altera o mapa da densidade de energia e a consequente refrigeração do DATACENTER. As cargas de trabalho passam a se movimentar dinamicamente com a VIRTUALIZAÇÃO e alteram o perfil energético do DATACENTER.

7.4.3. Conceito na Prática: VMware vFabric CLOUD Application Platform

A VMware lançou recentemente a plataforma de aplicação VMware vFabric CLOUD. A proposta é facilitar a construção de aplicações construídas baseadas em VIRTUALIZAÇÃO e CLOUD COMPUTING. A plataforma da VMware trabalha junto com o vSphere. Os principais aspectos da proposta da VMware são:

- Fornecer um framework para acelerar o desenvolvimento de aplicações.
- Fornecer uma plataforma de *runtime* otimizada para o framework e a infraestrutura de VIRTUALIZAÇÃO.
- Fornecer um conjunto de serviços de RUNTIME adaptados às necessidades de modernas aplicações.

O VMware Fabric inclui o framework Spring, usado por desenvolvedores para construir aplicações Java. O Spring permite a criação de diversos tipos de aplicações incluindo Java Enterprise, Web e integração Enterprise. Também vFabric incorpora o Apache TomCat, um servidor de *runtime* leve e já utilizado por muitas empresas.

7.5. CLUSTERIZAÇÃO

7.5.1. Introdução

Os *clusters* são parte importante na nova arquitetura de DATACENTERS. A VIRTUALIZAÇÃO resolve o problema da sobra de recursos em servidores físicos. A clusterização resolve o problema da falta de recursos em servidores físicos. Idealmente o *cluster* deve permitir que uma aplicação rode em mais de uma máquina física, incrementando a performance ao mesmo tempo em que aumenta a disponibilidade.

Os *clusters* podem ser classificados em:

- **Clusters de alta disponibilidade (High Availability – HA):** os *clusters* de alta disponibilidade endereçam redundância com capacidade de *failover* automático.
- **Clusters de balanceamento de carga (Load Balancing – LB):** os *clusters* do tipo *load balancing* endereçam a melhoria da capacidade para a execução da carga de trabalho (*WORKLOAD*).
- **Clusters de alta performance (HPC e HTC):** os *clusters* do tipo alta performance endereçam o aumento da performance da aplicação.

- **Clusters em Grid:** os *clusters* em grid endereçam o melhor dos mundos com alta disponibilidade e alto desempenho. Podem ser globais, empresariais ou simplesmente *grid clusters*.

7.5.2. Clusters de Balanceamento de Carga e de Alta Disponibilidade

A escolha da tecnologia de *cluster* HA ou LB depende se os aplicativos a serem executados possuem estado de execução demorada na memória. As diferenças entre os dois tipos de *clusters* são que:

- O *cluster* HA destina-se aos aplicativos que têm estado de execução demorada na memória ou que têm estados de dados frequentemente atualizados. Esses são denominados aplicativos com monitoração de estado e incluem aplicativos de bancos de dados e aplicativos de mensagens. A utilização típica dos *clusters* de *failover* inclui servidores de arquivos, servidores de impressão, servidores de bancos de dados e servidores de mensagens.
- O *cluster* LB destina-se aos aplicativos que não têm estado de execução demorada na memória. Esses são denominados aplicativos sem monitoração de estado. Um aplicativo sem monitoração de estado trata cada solicitação do cliente como uma operação independente e, portanto, pode balancear a carga de cada solicitação de forma independente. Em geral, os aplicativos sem monitoração de estado possuem dados somente de leitura ou dados frequentemente alterados. Servidores web, VPNs, servidores FTP, *firewalls* e servidores *proxy* costumam usar o *cluster* LB.

7.5.3. Cluster de Alta Performance

Uma forma de classificar os *clusters* de alto desempenho é pelo ambiente em que estão inseridos. Neste caso, os *clusters* podem ser de dois tipos:

- **HTC (*High Throughput Computing*):** foco no *throughput*. Exemplo de utilização é na área de finanças.
- **HPC (*High Performance Computing*):** foco no desempenho e é o mais comum. Exemplo de utilização é na área de meteorologia.

A **Figura 7-7** ilustra a arquitetura do DATACENTER voltado para alta performance.

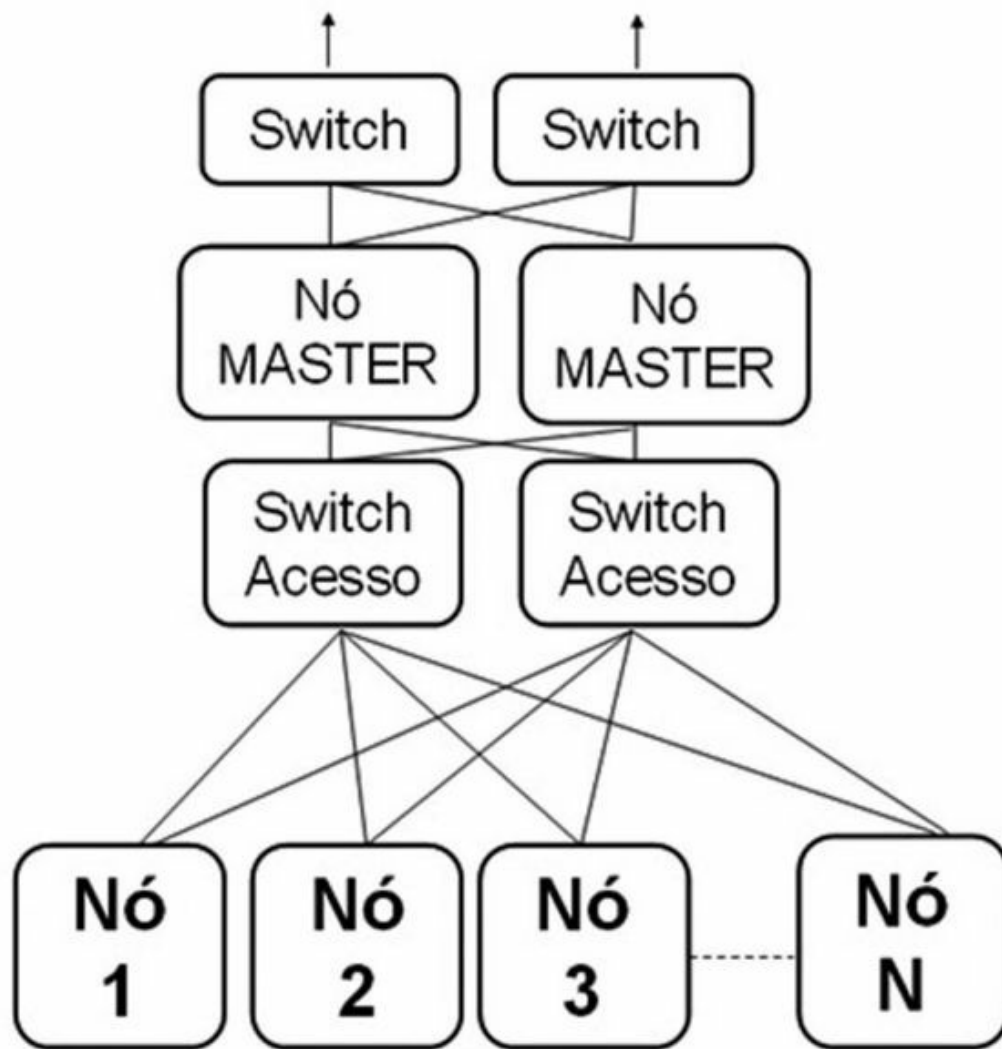


Figura 7-7 – Arquitetura do DATACENTER baseado em *Cluster*

Normalmente os *clusters* HPC e HTC são utilizados por universidades e centros de pesquisa que buscam construir sistemas de processamento de alta performance utilizando tecnologias abertas e de baixo custo. Boa parte dos *clusters* HPC em uso atualmente é baseada em arquiteturas x86 (AMD ou Intel) e utiliza os sistemas operacionais Windows ou Linux.

O *cluster* HPC é o mais comum de ser encontrado. Sua principal característica é o processamento de alto desempenho. Além de fornecer alto desempenho, este tipo de *cluster* também apresenta alta disponibilidade devido à maneira que é construído e ao suporte do sistema operacional.

Para o *cluster* HPC funcionar é necessário utilizar um software de gerenciamento que permita que os diversos nós de processamento existentes sejam tratados como um único nó. O Beowulf é o software mais conhecido utilizado no gerenciamento do HPC desenvolvido para o ambiente Linux.

7.5.4. Top 500 Supercomputers Site

A lista *Top 500 Supercomputers Site* publicada em www.top500.org fornece uma classificação com as 500 maiores arquiteturas de computação, quer utilizem *clusters* ou não. Os *clusters* HPC (*High Parallel Clustering*) representam 55% dos *clusters* existentes na última lista. Supercomputadores que utilizam a arquitetura de memória NUMA (*Non-Uniform Memory Access*)

também são muito comuns nesta lista.

A lista *Top 500* tem atualização de forma semestral e a última atualização é de junho de 2011 (trigésima sétima lista). O número 1 desta lista de jun/2011 é o *K computer* da *RIKEN Advanced Institute for Computational Science* (AICS), baseado no processador SPARC64 VIIIx 2000 MHz (16 *gigaflops*) e no sistema operacional Linux. Este *cluster* está no Japão e consome cerca de 10MW. A **Figura 7-8** ilustra o *K computer*.



Figura 7-8 – *K computer*

Os principais fornecedores dos computadores utilizados nos *clusters* HPC são a HP, que lidera a lista com 42% dos sites de supercomputadores da lista, seguida pelos 37% da IBM. O Linux é o sistema operacional mais utilizado nestes computadores. A facilidade de criação e a administração dos *clusters* com este sistema operacional, além do desempenho, estabilidade e relativa facilidade de uso, tornam-no uma boa escolha. Sua participação nesta lista está próxima a 85,4%. O Windows está em 28,6% dos supercomputadores, mas vale ressaltar que foi lançado depois.

7.5.5. Clusters em Grid

Os *grids* são uma evolução natural do conceito de DATACENTER em *cluster*. Os *clusters* são um aglomerado de máquinas conectadas em rede que executam serviços de rede. Os *grids* podem ser considerados uma organização virtual que permite a aglomeração de recursos distantes geograficamente. Uma espécie de evolução do conceito de *cluster*.

Assim como os DATACENTERS em *clusters*, os *grids* estão se tornando populares. A ideia

por trás tanto dos *clusters* quanto dos *grids* é basicamente a mesma: combinar o poder de processamento de vários computadores ligados em rede para conseguir executar tarefas que não seria possível (ou pelo menos não com um desempenho satisfatório) executar utilizando um único computador e ao mesmo tempo fazê-lo a um custo mais baixo do que o de um supercomputador de potência semelhante.

Os *grids* podem ser temporários, formados para executar uma tarefa específica e depois desfeitos. Presumindo que todos os computadores estejam previamente ligados em rede, criar e desfazer um *grid* é apenas questão de ativar e depois desativar o software responsável em cada computador.

Para alguns autores, a principal diferença entre um *cluster* e um *grid* é que um *cluster* possui um controlador central, um único ponto de onde é possível utilizar todo o poder de processamento do *cluster*. Os demais nós são apenas escravos que servem a este nó central.

Os *grids*, por sua vez, utilizam uma arquitetura mais flexível onde, embora possa existir algum tipo de controle central, tem-se um ambiente cooperativo, onde os usuários compartilham os ciclos ociosos de processamento em seus sistemas em troca de poder utilizar parte do tempo de processamento do *grid*. É uma ideia mais inteligente, mas também é mais complexa.

Os *grids* podem ter uma dimensão global, empresarial ou se confundir com um *cluster*. Em *grid*, duas empresas sediadas em países com fusos-horário diferentes poderiam formar um *grid*, combinando seus servidores, de modo que uma possa utilizar os ciclos de processamento ociosos da outra em seus horários de pico, já que com horários diferentes os picos de acessos aos servidores de cada empresa ocorrerão em horários diferentes.

As nuvens e os *grids* compartilham visões similares, pois ambos reduzem os custos da computação, aumentando a flexibilidade e otimizando o uso dos recursos, mas apresentam inúmeras diferenças, passando pelo compartilhamento dos recursos (recursos não compartilhados na computação de nuvens versus recursos compartilhados no *grid*), o uso da VIRTUALIZAÇÃO (na computação de nuvens a VIRTUALIZAÇÃO é o elemento central, o que ainda não é uma realidade no *grid*), a forma de implementar a segurança (na computação de nuvens a segurança é obtida pela isolamento e, no *grid*, pelas credenciais e consequente delegação) e o grau de centralização (*grid* apresenta o controle descentralizado contra a nuvens que centraliza o controle).

7.5.6. Conceito na Prática: Oracle Exadata

O Oracle Exadata Database Machine é uma tecnologia de banco de dados que oferece alto desempenho tanto para *data warehousing* quanto para aplicações de processamento de transações on-line (OLTP), o que o torna uma boa escolha para consolidação em grades ou nuvens privadas. Trata-se de um pacote de servidores, armazenamento, rede e software com escalabilidade, segurança e redundância. Com o Oracle Exadata Database Machine, os clientes podem reduzir os custos de TI por meio de consolidação, gerenciar mais dados em várias camadas de compressão, melhorar o desempenho de todas as aplicações e tomar melhores decisões de negócios em tempo real.

O lançamento do Oracle Exadata Database Machine X2-8, uma solução completa com servidores de base de dados de oito soquetes, 14 Oracle Exadata Storage Servers e switches InfiniBand, estende a família de Exadata Database Machines, oferecendo aos clientes uma plataforma de consolidação excelente para aplicações OLTP e de *data warehousing* muito grandes.

O Oracle Exadata Database Machine X2-2 (antes conhecido como V2), versão lançada anteriormente, inclui oito servidores de banco de dados com dois soquetes, 14 Oracle Exadata Storage Servers e switches InfiniBand. Está disponível em diversas configurações que variam de um quarto de RACK a um RACK completo de 42 unidades.

Vantagens da solução Exadata:

- **Alto desempenho para *data warehouses*:** o Exadata Smart Scan melhora o desempenho de consulta transferindo o grande volume de processamento de consultas e pontuações de *data mining* para servidores de armazenamento escaláveis inteligentes.
- **Alto desempenho para aplicações OLTP:** o Exadata Smart Flash Cache faz cache de dados “quentes” em armazenamento de estado sólido rápido, de modo transparente, melhorando os tempos de respostas de consulta e a taxa de transferência.
- **Alto desempenho para cargas de trabalho consolidadas:** a enorme grade paralela do Exadata é ideal para consolidar data warehousing e aplicações OLTP, enquanto os componentes de gerenciamento de recursos de qualidade de serviço do Exadata asseguram melhores tempos de resposta para todos os usuários.

7.6. Blocos de Construção da TI

7.6.1. Introdução

A tendência atual é que os grandes blocos da TI formados por equipamentos de rede, servidores e *storage*, além da VIRTUALIZAÇÃO, sejam entregues em blocos pré-testados em configurações predefinidas, garantindo melhor interoperabilidade entre produtos de fabricantes diferentes.

A **Figura 7-9** ilustra a nova ideia.

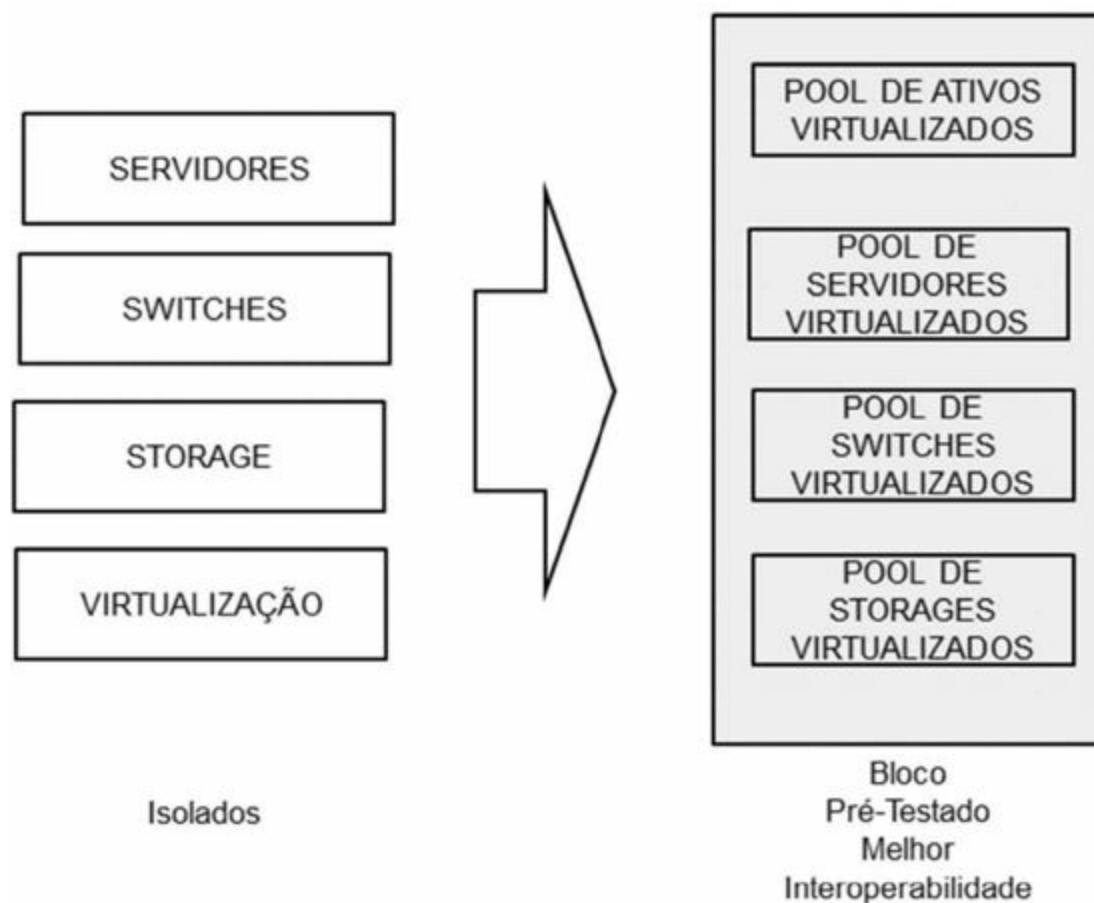


Figura 7-9 – Blocos de Construção da TI do DATACENTER

7.6.2. Conceito na Prática: VCE da VMware, Cisco e EMC

VMware, Cisco e EMC estão juntas em uma parceria chamada de VCE (Virtual Computing Environment) para oferecer uma solução para nuvens privadas baseadas no conceito de vBlock. Os vBlocks são pacotes de infraestrutura otimizados para a construção da nuvem privada.

Estas estruturas integradas facilitam a construção dos grandes blocos da nuvem privada e ofertam de maneira integrada: VIRTUALIZAÇÃO, processamento, armazenamento, rede, segurança e gerenciamento integrado.

Os pacotes vBlock são pré-testados e validados com desempenho, capacidade e disponibilidade definidos. A promessa é disponibilizar um caminho acelerado para a formação de nuvens privadas.

Os pacotes vBlock surgiram de uma ideia de simplificar a aquisição, a distribuição e a operação da infraestrutura de TI. O valor da solução encontra-se em uma combinação da eficiência, controle e escolha. Outro princípio que guia a ideia dos pacotes de infraestrutura como o vBlock é a habilidade de expandir a capacidade dos pacotes porque a arquitetura é muito flexível.

Os pacotes vBlock são otimizados para fornecimento de SLAs padrões, reduzem o risco e atendem necessidades de *compliance*. Para os clientes, simplificam a expansão e a escalabilidade, permitem adicionar *storage* e capacidade quando requeridos, podem se conectar à infraestrutura existente de LAN e permitem administração *multitenant* e segurança baseada em regras.

Os pacotes vBlock definem as necessidades de espaço e peso, energia e refrigeração e

predeterminam a escalabilidade e capacidade do DATACENTER.

Existem três categorias de pacotes do tipo vBlock:

- **vBlock 2:** projetado para instalações high-end. Envolve Cisco Unified Computing System, EMC Symmetrix VMAX *storage* e VMware vSphere 4.
- **vBlock 1:** projetado para grandes instalações. Envolve Cisco Unified Computing System, EMC Clarrion CX4 ou EMC Celerra unified *storage* e VMware vSphere.
- **vBlock 0:** projetado para sites moderados e grande número de máquinas virtuais. Envolve Cisco Unified Computing System, EMC Celerra unified *storage* e VMware vSphere 4.

A **Tabela 7-1** resume as características principais dos três blocos.

Tabela 7-1 – Configurações do vBlock

	vBlock 0	vBlock 1	vBlock 2
HYPERVISOR	vSphere 4 ESXi	vSphere 4 ESX	vSphere 4 ESX
Método de BOOT	Local	SAN	SAN
vCenter	Sim	Sim	Sim
Nexus 1000v	Sim	Sim	Sim
UCB B-series	UCB B-200/250	UCB B-200/250	UCB B-200/250
UCB Mínimo	4*B200, Intel X5550, 48GB RAM	12*B200, Intel X5570, 48GB RAM	32*B200, Intel X5570, 96GB RAM
UCB Máximo	16*B200, Intel X5550, 48GB RAM	24*B200, Intel X5570, 48GB RAM	64*B200, Intel X5570, 96GB RAM
CNA (FCoE)	Emulex, Qlogic, VIC	Emulex, Qlogic, VIC	Emulex, Qlogic, VIC
VMs	128-512	512-3000	3000-20000
UCS Manager	Sim	Sim	Sim
Storage	CELERRA NS-120	CLARIION CX4-480	SYMMMETRIX V-MAX
Storage Mínimo	5*600GB 15K 5*1TB SATA	6*400GB EFD 78*450GB 15K 21*1TB SATA	9*400GB EFD 124*450GB 15K 76*1TB SATA
Storage Máximo	25*600GB 15K 15*1TB SATA	16*400GB EFD 78*450GB 15K 146*450GB 15K	21*1TB 27*1TB SATA
NAS GATEWAY	N/A	NS-G2 (recomendado)	NS-G8 (recomendado)
Switch SAN	N/A	MDS 9506 (preferido)	MDS 9506 (preferido)
UIM	Preferível	Preferível	Preferível

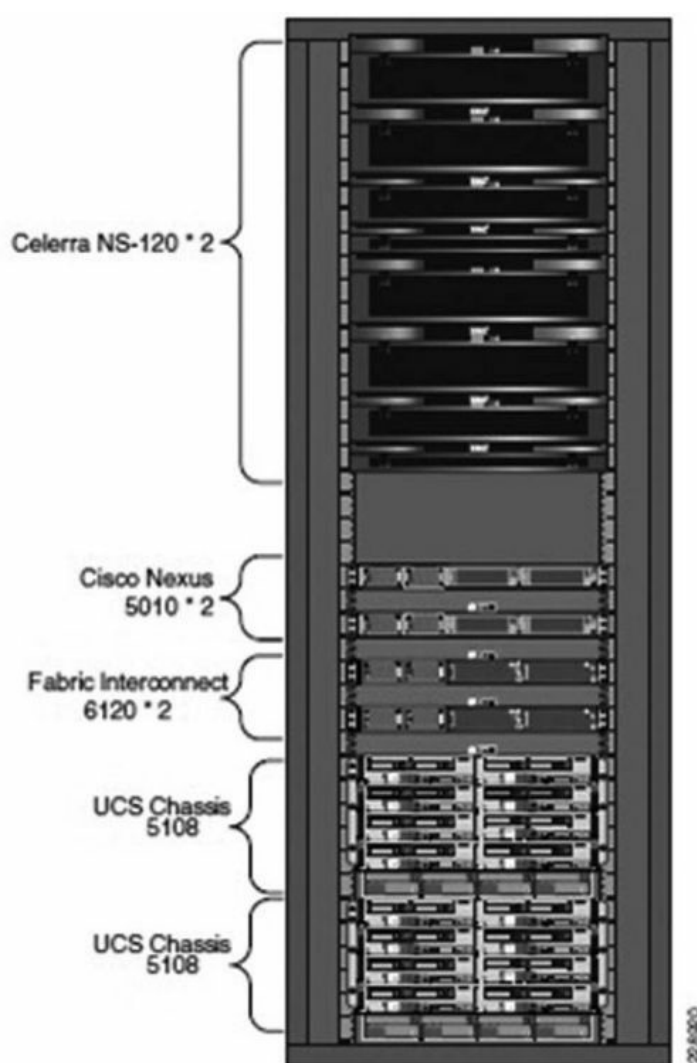


Figura 7-10 – vBlock RACK LAYOUT

EMC Ionix

O UIM (*Unified Infrastructure Manager*, gerenciador de infraestrutura unificada), do EMC Ionix, permite provisionamento simplificado e integrado, gerenciamento de configuração, alteração e conformidade para VIPs (*vBlock Infrastructure Packages*, pacotes de infraestrutura vBlock).

O UIM foi criado especificamente para o gerenciamento de elementos da infraestrutura vBlock, gerenciando os vBlocks como uma só entidade. Isso reduz as despesas operacionais – enquanto fornece recursos de gerenciamento simplificado que capacitam as empresas a fazer mais facilmente a transição de infraestruturas físicas para virtuais e para *CLOUD* privada. Na versão inicial, o UIM oferecerá gerenciamento de suporte total do UCS (*Unified Computing System*, sistema de computação unificada) da Cisco. Em seguida, evoluirá para um suporte completo das infraestruturas vBlock.

Os principais recursos do UIM incluem :

- **Painel de controle vBlock e portal de provisionamento de TI:** com o painel de controle do UIM, é possível visualizar diversas implantações de vBlock, dando a você exibições consolidadas de toda a sua infraestrutura vBlock.

- **Catálogo dos perfis de serviços:** perfis de serviços são a “receita” para a criação de serviços e a base para a entrega da “infraestrutura como serviço”. Com o UIM, é possível criar vários níveis de perfis de serviços para configurações de rede, computação e armazenamento e combiná-los em uma “oferta” para os clientes. Isso habilita os componentes vBlock a serem fornecidos como unidades padrão de serviço para um cliente, selecionado por meio do painel de controle vBlock. Os recursos do perfil de serviços serão fornecidos para a infraestrutura Cisco na versão inicial.
- **Gerenciamento baseado em políticas:** as políticas de configuração podem ser definidas e aplicadas para garantir a conformidade em todo o sistema, a fim de evitar um fluxo de configurações. A primeira versão do UIM do Ionix oferece esses recursos para o UCS da Cisco e para a infraestrutura de rede relacionada, como o Cisco Nexus e dispositivos MDS.
- **Provisionamento unificado, configuração e alteração:** o UIM oferece uma implantação de infraestrutura automatizada e provisionamento de hardware vazio. Isso inclui um provisionamento automatizado de infraestrutura de local para recuperação de desastres. O UIM aproveita o contexto entre domínios para gerenciar dependências entre software, hardware, rede e armazenamento. A detecção profunda e granular oferece uma base para o provisionamento, a configuração e a alteração do UIM – permitindo um histórico de revisão ilimitado e rastreamento individualizado, capacidade de rastreamento e de reprodução. A primeira versão do UIM concentra-se no provisionamento unificado, na configuração e na alteração para o UCS da Cisco e para a infraestrutura de rede relacionada, como o Cisco Nexus e dispositivos MDS.
- **Integração simplificada:** o gerenciamento de elementos vBlock UIM integra-se a sua solução de gerenciamento corporativo, oferecendo eventos de alteração e conformidade para ajudá-lo a rastrear alterações essenciais. O UIM também aproveita as APIs do UCS Manager da Cisco e os sistemas de gerenciamento de armazenamento da EMC, obtendo os benefícios da instrumentação existente e da configuração de componentes.

Recentemente a Cisco, a VMware e a Network Appliance lançaram o FLEXPOD, que é um bloco semelhante ao VCE. Maiores informações podem ser encontradas nos sites dos respectivos fabricantes.

7.6.3. Padrão FCoE

Os padrões e tecnologias das redes de *storage* vêm sendo alterados ao longo dos últimos quinze anos. A utilização de um único padrão é um sonho. Na prática as organizações já possuem redes de armazenamento e padrões estabelecidos e novos padrões estão chegando. O segredo do FCoE é integrar novas opções com arranjos já instalados e maduros, evitando o rompimento tecnológico precoce. Esta é a proposta do padrão FCoE.

Os padrões de protocolo utilizados nas redes de armazenamento podem ser resumidos conforme descrito a seguir:

- Fibre Channel (FC) é protocolo de armazenamento de bloco padrão para redes de *storage* (SAN). Em DATACENTERS corporativos aproximadamente 90% dos dispositivos de

armazenamento são conectados via FC. A SAN permite a consolidação do *storage*, gerenciamento centralizado, alta performance e confiabilidade e reconfiguração rápida. Mesmo assim FC não conseguiu ser o padrão para *fabrics* de *clusters* e redes de volume devido à vantagem do padrão Ethernet em termos de custo. Mesmo indo para 8Gb/s, o padrão FC será mais lento do que o padrão Ethernet, que deve se mover rapidamente para 40GB/s e 100Gb/s nos próximos anos.

- iSCSI transporta SCSI over TCP/IP e possibilita que SANs sejam criadas utilizando switches Ethernet de baixo custo e softwares *iSCSI initiators*. Tem sido utilizado como um protocolo para blocos em redes SMB e agora também com adaptadores Ethernet em 10 Gigabits utilizando TOE (TCP Offload Engines).
- SAS (*Serial Attached SCSI*) é uma alternativa focada em custo que permite trocar o SCSI paralelo. Mas apresenta limites de escalabilidade, performance e distância quando comparada com FC em redes de *storage*.
- NAS *appliances* utilizam protocolos NFS e CIFS para sistemas de arquivos acessados através de redes Ethernet.

Por que então utilizar mais um padrão? A ideia do padrão FCoE é transportar os frames FC através de um *fabric* Ethernet, possibilitando a melhoria da performance da aplicação enquanto reduz custo, energia e gerenciamento convergindo *storage*, rede e *clusters* em um único *fabric*.

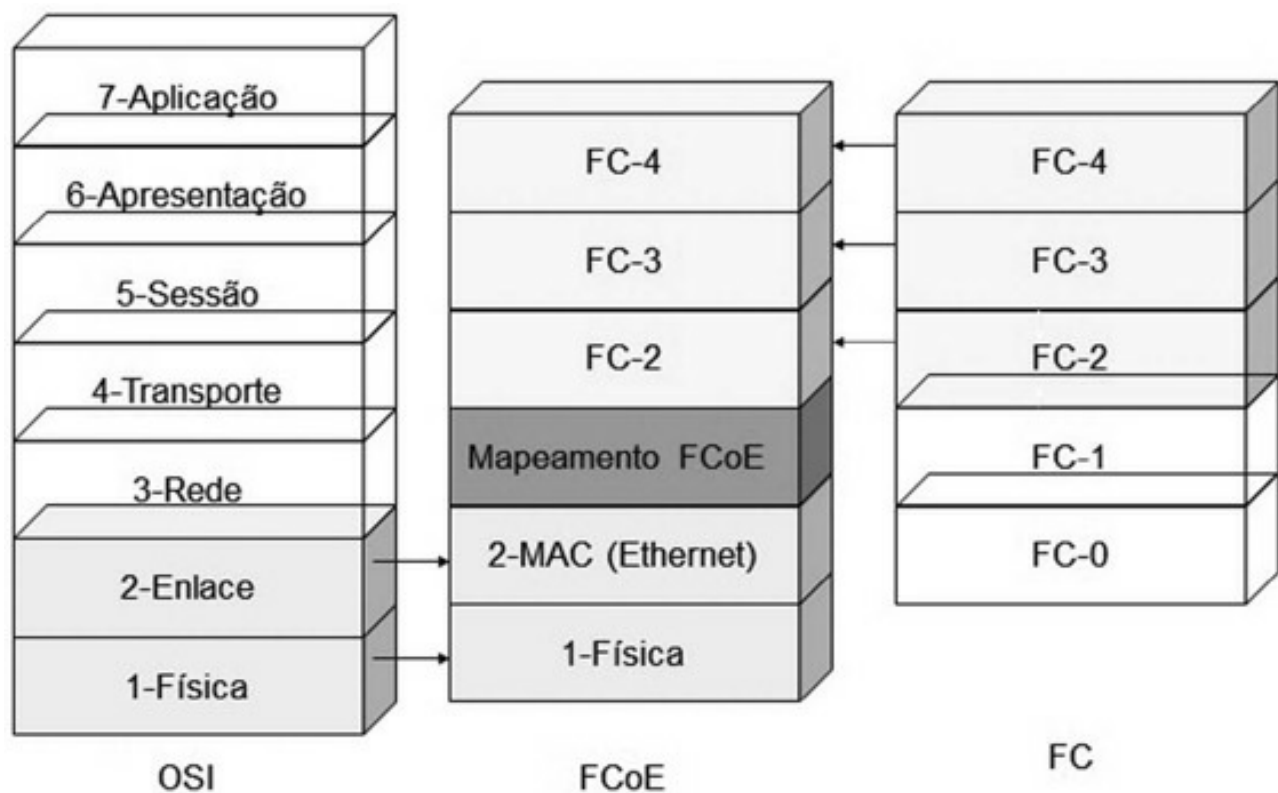


Figura 7-11 – Mapeamento Fibre Channel para Ethernet

Os novos switches baseados em FCoE trazem o conceito inovador de *unified fabric*, conjunto de tecnologias que permitem a consolidação das diversas redes existentes em um DATACENTER

com uma única infraestrutura de rede.

Com a tecnologia *unified fabric*, as diversas redes Ethernet (LAN), redes de armazenamento Fibre Channel (SAN) e redes de computação de alto desempenho (HPC) podem ser implementadas sobre uma infraestrutura Ethernet unificada, permitindo maior flexibilidade, compartilhamento de recursos e redução de custos com cabeamento, energia e equipamentos dedicados.

No caso de *blades*, a situação é ainda mais crítica, pois a exigência de mais portas de I/O para maior número de cores por processador e maior uso de máquinas virtuais reforça a necessidade de integração de tecnologias.

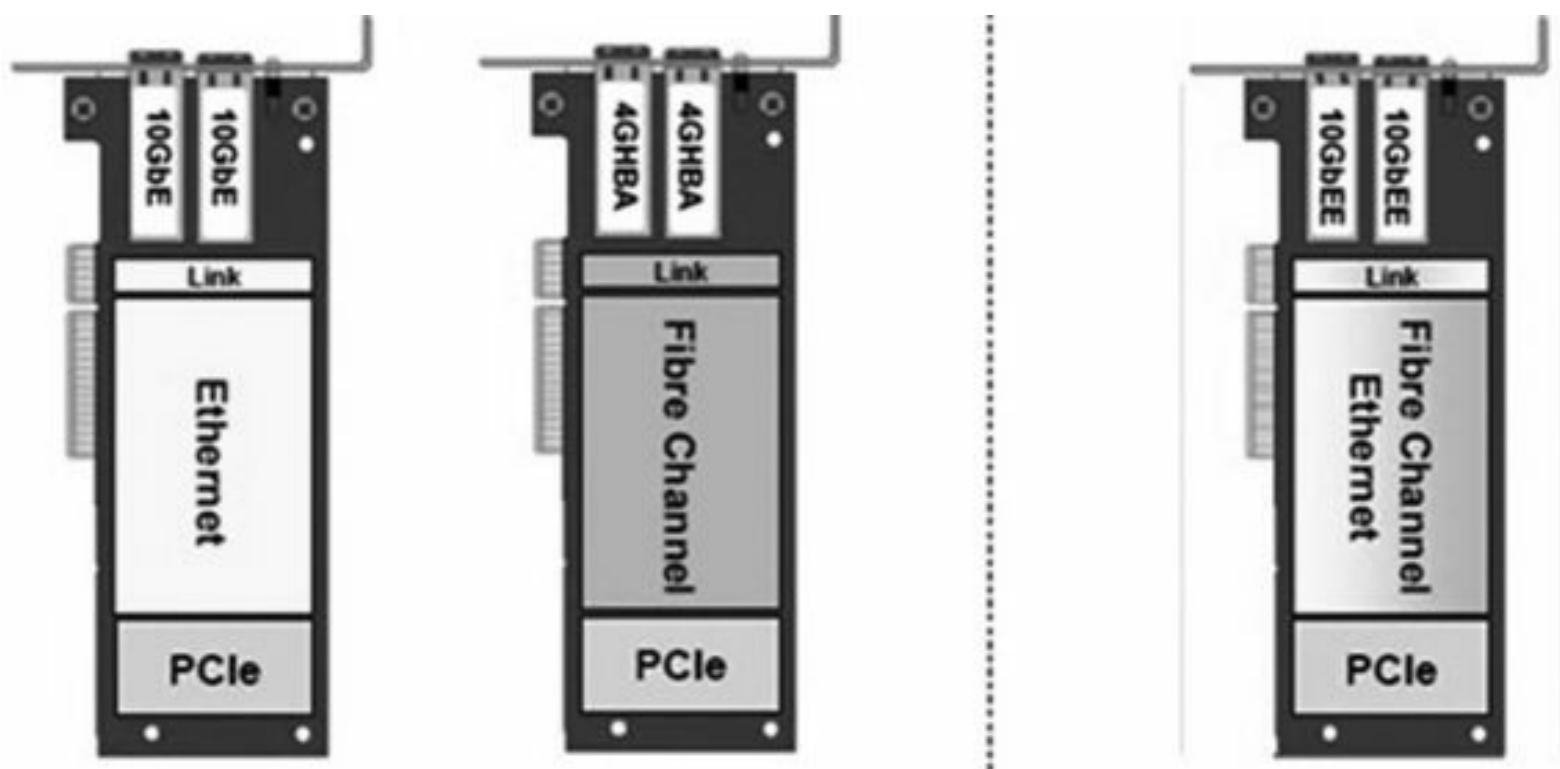
O *unified fabric* reduz o custo eliminando a necessidade de utilizar múltiplos adaptadores para dados específicos como FC para *storage* e Infiniband para *clustering*. Enquanto o padrão 10 Gbps permite obter alta performance com o FCoE, o FCoE também pode rodar em redes existentes FC sem custos adicionais. O projeto Open-FC.org está criando os drivers para o SO Linux.

Os switches unificados adotam o padrão DATACENTER Ethernet com melhorias no controle de fluxo e gerenciamento do congestionamento da rede (padrão 802.1p QoS). Os switches encapsulam o frame FC no pacote Ethernet (FCoE), suportando a consolidação do I/O.

Assim, o padrão FCoE permite conduzir o tráfego Fibre Channel através de uma rede Ethernet. A **Figura 7-12** ilustra a ideia de redução dos adaptadores no padrão FCoE. Substituem-se a LAN e a HBA por um cartão CNA (adaptador de rede convergente – *Converged Network Adapter*).

A unificação das redes SAN e LAN decorre das tecnologias Ethernet e FCoE (*Fibre Channel over Ethernet*), onde a rede Ethernet passa a ser o novo meio físico das redes lógicas LAN e SAN. Os servidores possuem acesso à LAN e à SAN pelo adaptador CNA, eliminando a necessidade de NIC para LAN e HBA ou outra NIC para SAN.

Os resultados da convergência LAN/SAN são: custos menores com infraestrutura de rede, cabeamento, energia, resfriamento, gerenciamento, além da consolidação em si.



O padrão FCoE aos poucos vem se consolidando como o padrão para interligação das partes do DATACENTER. A grande vantagem deste padrão é que, ao mesmo tempo em que ele simplifica a conexão das redes de *storage*, ele não abre mão do desempenho. A simplificação trazida pelo FCoE pode ser mostrada na **Figura 7-13** onde as linhas cheias indicam conexão Ethernet, linhas tracejadas indicam conexão FC, linhas cheias e tracejadas indicam conexão FCoE.

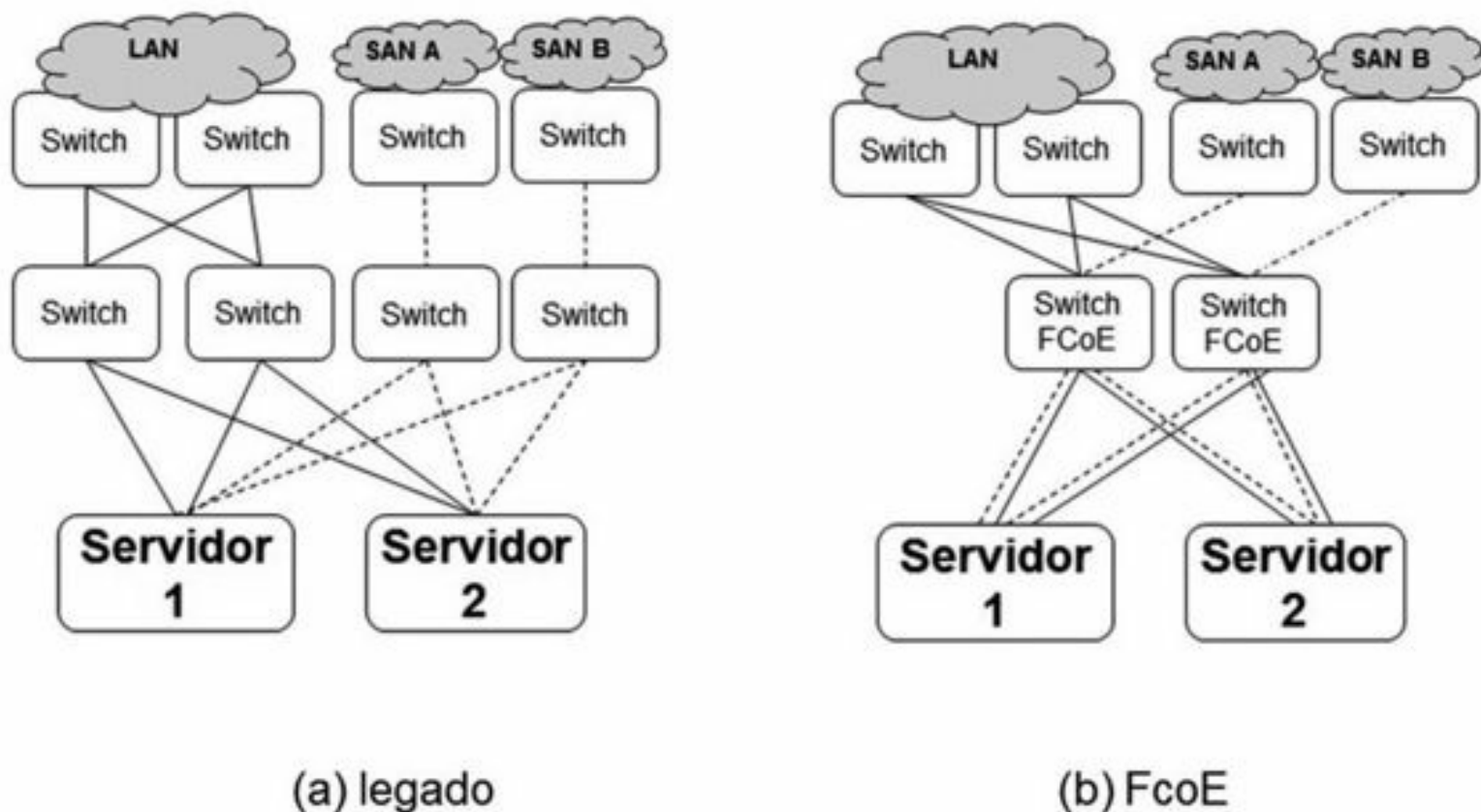


Figura 7-13 – Integração com e sem FCoE

7.6.4. Conceito na Prática: Open FCoE da Intel

A Intel deu um passo para simplificar os DATACENTERS ao apresentar o Open FCoE, que permite que todo o tráfego do DATACENTER rode em um único cabo usando a família de adaptadores para servidores Intel 10 Gigabit Ethernet (10 GbE) X520. O padrão FCoE aos poucos está se consolidando. O Open FCoE da Intel ajuda ainda mais e deve tornar o FCoE o padrão para conexão SAN e LAN.

Redes unificadas permitem que os departamentos de TI criem conexões flexíveis em DATACENTERS virtualizados ao consolidar múltiplas redes de dados e armazenamento em uma única rede 10GbE. A consolidação pode ajudar a reduzir o gasto mundial com TI em US\$ 3 bilhões por ano. Os 122 mil quilômetros de cabos dos DATACENTERS globais economizados são suficientes para enrolar a Terra três vezes.

Uma rede unificada simples e de alta velocidade para o DATACENTER é um dos pilares da

visão “CLOUD 2015” da Intel e da Open Data Center Alliance, anunciada em outubro de 2010. Uma rede unificada com 10 GbE cria uma infraestrutura de dados mais simples, que é mais fácil de gerenciar e ainda assim pode acomodar o intenso tráfego de rede da nuvem.

O Open FCoE da Intel integra capacidades ao sistema operacional para fornecer uma rede unificada sem a necessidade de hardware proprietário adicional. Os departamentos de TI podem usar ferramentas de gestão comuns para redes de servidores e conectividade de armazenamento ao mesmo tempo em que se integram com os ambientes de Fibre Channel. O Open FCoE é diferente das soluções baseadas em HBA FCoE (CNA), pois parte do *stack* FC e o FCoE *transport protocol* está integrada ao próprio SO (*initiator* está integrado ao SO), simplificando o *device driver*. Adaptadores convencionais em 10 GbE estão agora prontos para a SAN.

Uma estrutura unificada oferece suporte tanto para recursos computacionais quanto para recursos de armazenamento, por meio de um transporte com alta largura de banda para oferecer uma melhor eficiência para os DATACENTERS, além de simplificar a gestão e poder acelerar a implementação de serviços baseados em nuvens e de VIRTUALIZAÇÃO. Os switches Cisco Nexus 10 Gigabit Ethernet e os servidores Cisco Unified Computing System oferecem suporte para os adaptadores Open FCoE 10 Gigabit Ethernet da Intel para fornecer mais opções para o acesso unificado, expansível e com melhor custo-benefício. As empresas que estão apoiando a solução Open FCoE, além da Intel, incluem Cisco, Dell, EMC, NetApp, Oracle e Red Hat. A **Figura 7-14** ilustra o adaptador FCoE da Intel.

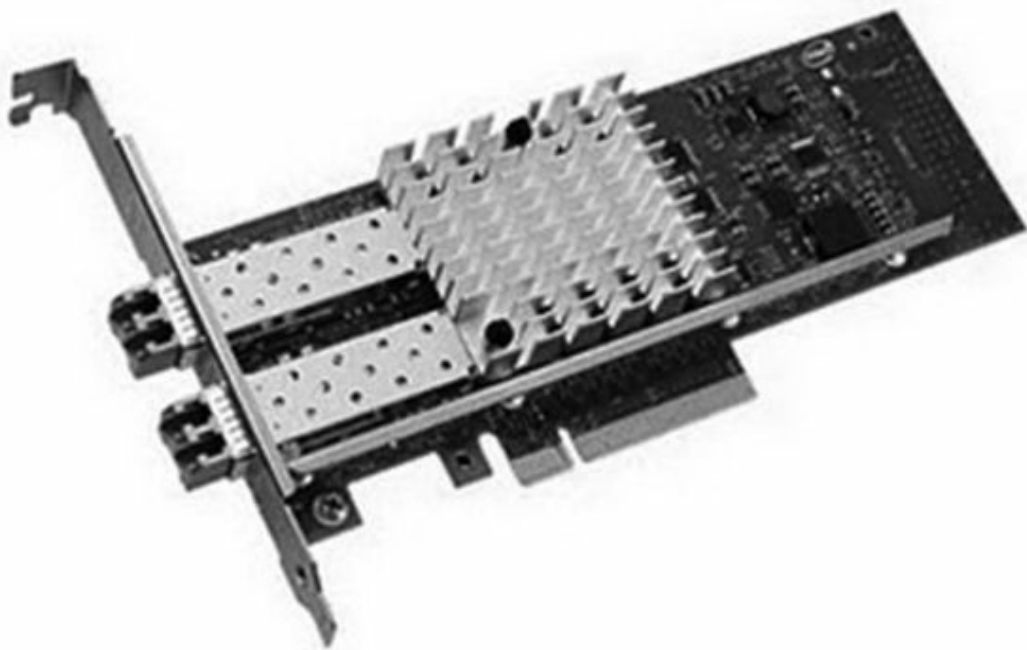


Figura 7-14 – Adaptador 10 GbE FCoE da Intel

7.7. Referências Bibliográficas

10 Gigabit Ethernet Unifying Fabric: Foundation for the Scalable Enterprise, Dell Power Solutions, Server Virtualization: Cisco Unified Computing System, Cisco Systems, 2009.
Blades Platforms and Network Convergence, [Blade.org](http://blade.org), 2008.

- Cisco DATACENTER 3.0. At a Glance. Cisco Systems, 2009.
- Cisco UCS Management. At a Glance, Cisco Systems, 2010.
- Cisco Unified Computing System. At a Glance. Cisco Systems, 2010.
- Cisco VN-Link Virtualization-Aware Networking, Cisco Systems, 2009.
- Clark, Tom. Designing Storage Area Network: A practical reference for Implementing Fibre Channel and IP SANs, Addison-Wesley, Second Edition, 2003.
- Clark, Tom. IP SANs: A guide to iSCSI, iFCP and FCIP Protocols for Storage Area Networks, Addison Wesley, 2004.
- DATACENTER Blade Server Integration Guide, Cisco, 2006.
- Dell Business Ready Configuration for DATACENTER virtualization, Dell, 2009.
- Fabric convergence with lossless Ethernet and Fibre Channel over Ethernet (FCoE), Technology Bulletin, Blade Network Technologies, 2008.
- Information Storage and Management: Storing, Managing, and protecting Digital Information, EMC Education Services, 2009.
- Veras, Manoel. DATACENTER: Componente Central da Infraestrutura de TI, Brasport, 2009.
- Veras, Manoel. VIRTUALIZAÇÃO: Componente Central do DATACENTER, Brasport, 2011.

PARTE III: ***SERVIÇOS DE NUVEM***

8. Infraestrutura como Serviço (IaaS)

8.1. Introdução

Infraestrutura como um serviço (*Infrastructure as a Service – IaaS*) é a capacidade que o provedor tem de oferecer uma infraestrutura de processamento e armazenamento de forma transparente para o cliente, normalmente uma organização. Neste cenário, os usuários da organização não têm o controle da infraestrutura física, mas, através de mecanismos de VIRTUALIZAÇÃO, possuem controle sobre as máquinas virtuais, armazenamento, aplicativos instalados e possivelmente um controle limitado dos recursos de rede. Um exemplo de IaaS é a opção Amazon EC2.

Infraestrutura como Serviço é basicamente deixar de adquirir hardware e software básico e passar a desenvolver a aplicação em uma infraestrutura virtual baseada na Internet e adquirida e paga na forma de serviço.

8.2. Conceito na Prática: Amazon AWS

8.2.1. Introdução

Os serviços Amazon Web Services (AWS) fornecidos pela Amazon são o que há de mais próximo na atualidade do modelo teórico proposto pelo NIST. A Amazon vem desenvolvendo estes serviços de nuvem desde 2006 e hoje possui uma oferta única de serviços com foco em IaaS. A Amazon aproveitou sua experiência no suporte a sua plataforma de e-commerce e criou o Amazon AWS.

Os serviços AWS permitem o acesso a recursos de computação, armazenamento e banco de dados e outros serviços de infraestrutura *on demand*. A ideia é que esta forma de computação reduza custos, melhore o fluxo de caixa da organização contratante, minimize os riscos do negócio e maximize as oportunidades. Segundo a Amazon, o serviço tradicional de DATACENTER, o hosting, já explicado anteriormente, quando comparado ao serviço de nuvem, é pouco eficaz, considerando que o uso e a capacidade dos recursos estão otimizados no modelo AWS.

A plataforma da Amazon se propõe a ser uma plataforma com pouca interação humana no que diz respeito ao suporte das aplicações. Ou seja, a ideia é que a plataforma funcione de preferência sem intervenção humana, melhorando a eficiência.

A Amazon permite obter uma boa flexibilidade para o desenvolvimento de aplicativos, pois as empresas podem continuar a usar o modelo de programação, sistemas operacionais, banco de dados e arquiteturas que já são familiares. As empresas podem inclusive misturar arquiteturas para servir aos diferentes modelos de negócio. Com o AWS as organizações podem subtrair e somar recursos

para suas aplicações para atender as demandas dos clientes e custos de gerenciamento.

É lógico que a forma de programar novas aplicações deveria quase que naturalmente utilizar os novos recursos disponibilizados pela plataforma AWS, que permite a construção de aplicações tolerante, a falhas e escaláveis, mas esta não é uma tarefa fácil, considerando que a empresa normalmente já esta operacional, e migrar aplicativos alterando suas arquiteturas não é simples.

O AWS também fornece privacidade de dados com a encriptação e publica procedimentos de backup e redundância.

A **Figura 8-1** ilustra a infraestrutura montada pela Amazon para os serviços AWS.

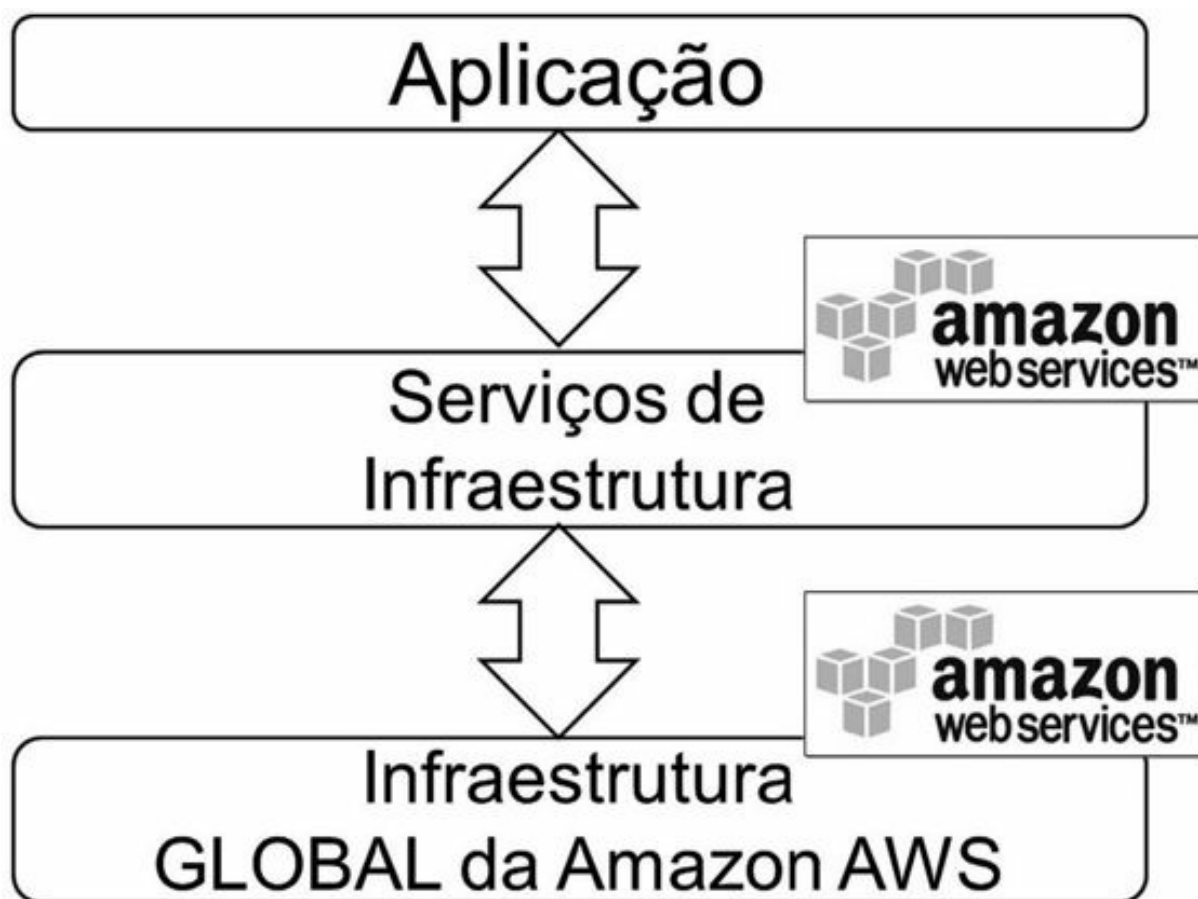


Figura 8-1 – Serviços AWS da Amazon

A proposta dos serviços AWS é fornecer serviços do tipo IaaS com:

- Flexibilidade.
- Efetividade.
- Escalabilidade.
- Elasticidade.
- Segurança.

Os elementos centrais dos serviços de infraestrutura da Amazon que fornecem os blocos de construção mais comuns necessários às aplicações são:

- **Computação:** o Amazon Elastic Compute Cloud (EC2) fornece a capacidade de aumentar ou diminuir recursos de computação baseando-se na demanda e facilita o fornecimento de novas instâncias de servidor. O EC2 é a máquina virtual.
- **Armazenamento:** todos necessitam de armazenamento – para arquivos, documentos, downloads de usuários ou backups. É possível armazenar qualquer coisa que o aplicativo necessite no Amazon Simple Storage Service (S3) e tirar vantagem do armazenamento escalável, confiável, altamente disponível e de baixo custo.
- **Armazenamento de dados:** Amazon SimpleDB (SDB) fornece armazenamento escalável, indexado e sem manutenção, em conjunto com processamento e enfileiramento para conjuntos de dados.
- **Sistema de Mensagens:** permite desacoplar componentes das aplicações usando o sistema de mensagens fornecido pelo Amazon Simple Queue Service (SQS). O SQS permite construir aplicações escaláveis.

Há dois níveis de suporte disponíveis para usuários do Amazon Web Services:

- **AWS Basic Support.** Suporte baseado em fórum grátis da equipe da Amazon que monitora os fóruns Amazon.
- **AWS Premium Support.** Pacotes de suporte pagos que fornecem suporte individual e por telefone e são uma maneira mais discreta de solicitar ajuda.

A **Figura 8-2** ilustra as opções do AWS Premium Support.

	Bronze	Silver	Gold	Platinum
Access to community forums	✓	✓	✓	✓
Resolution for AWS-owned issues	✓	✓	✓	✓
Local business hours	✓	✓	✓	✓
One-on-one online support	✓	✓	✓	✓
Client side diagnostic tools	✓	✓	✓	✓
Best practice guidance	✓	✓	✓	✓
Fastest guaranteed response	12 hours	4 hours	1 hour	15 minutes
Your named contacts (what's this? ⓘ)	1	2	3	unlimited
Architecture Support (what's this? ⓘ)	Building Blocks	Service Reviews	Use Case Guidance	Application Architecture
Always available — any time, any day, 24/7/365			✓	✓
One-on-one phone support			✓	✓
Direct access to Technical Account Manager				✓
White-glove case routing (what's this? ⓘ)				✓
Management business reviews (what's this? ⓘ)				✓

Figura 8-2 – Opções do Premium Support do AWS

A Amazon publica o status de funcionamento de todos seus serviços da Web em um painel publicamente acessível que é atualizado com qualquer problema relacionado aos serviços. O endereço do status dos serviços é publicado em:

<http://status.aws.amazon.com/>

Durante a interrupção do serviço, a equipe do Amazon Web Services atualiza o status a cada 15-30 minutos, enquanto trabalha para solucionar o problema.

8.2.2. Serviços AWS

É possível combinar serviços AWS quando e como necessário estes foram desenvolvidos para trabalhar em conjunto.

Por estar sendo executada dentro de um ambiente Amazon, toda a comunicação entre esses serviços será, geralmente, muito rápida.

Os serviços AWS são ilustrados na **Figura 8-3**.

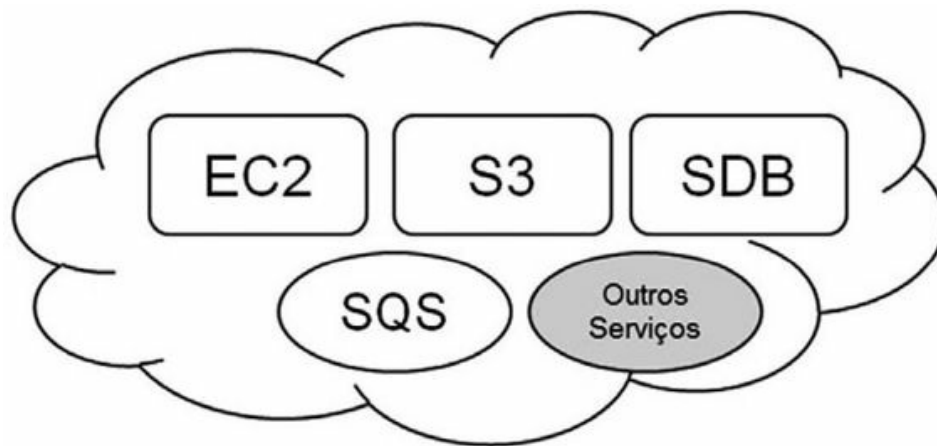


Figura 8-3 – Nuvem AWS

Amazon Elastic Compute Cloud (Amazon EC2)

O Amazon EC2 é um serviço web que permite aos desenvolvedores utilizar os recursos de nuvem sob demanda. O Amazon EC2 também permite pagar pelo uso. A criação de novas instâncias de servidor pode ser realizada em minutos. Também permite que as aplicações, desde que construídas adequadamente, estejam isoladas e livres de cenários de falha comuns em ambientes não controlados. A opção *Auto Scaling*, uma espécie de funcionalidade do EC2, permite que a capacidade EC2 cresça ou diminua de acordo com a demanda, minimizando os custos. O *Auto Scaling* é útil para aplicações que variam o uso semanalmente ou até diariamente. O *Auto Scaling* é operacionalizado pelo serviço Amazon CloudWatch.

No EC2 paga-se por instâncias com base no tipo de instância e no seu uso por hora real. Os servidores virtuais são executados dentro do ambiente seguro dos próprios DATACENTERS da Amazon.

O EC2 pode proporcionar aos aplicativos web a capacidade de:

- Configurar os requisitos de computação instantaneamente.
- Ajustar a capacidade com base na demanda.

Amazon Elastic Load Balancing, Auto Scaling, CloudWatch e os volumes EBS são funcionalidades do EC2, como ilustra a **Figura 8-4**.

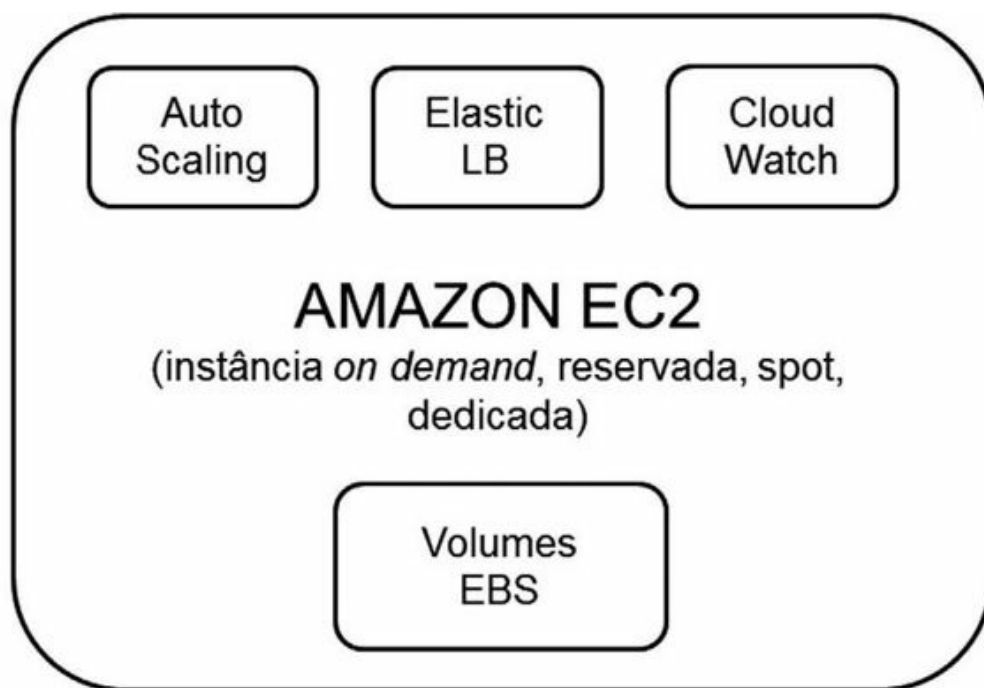


Figura 8-4 – Amazon EC2

- **Amazon Auto Scaling (Amazon AS):** o *auto scaling* automaticamente provisiona recursos computacionais do serviço Amazon EC2. No serviço AS, definem-se regras que determinam quanto mais (ou menos) instâncias de servidores são necessárias. As métricas são coletadas do serviço Amazon CloudWatch, que monitora as instâncias EC2. A utilização do Auto Scaling em conjunto com outros componentes da arquitetura AWS é mostrada na [Figura 8-5](#). Verifica-se que o AS trabalha em conjunto com o CloudWatch e o ELB para prover escalabilidade às aplicações. A escalabilidade pode ser baseada em eventos do tipo *scale-out* (instâncias são adicionadas) ou *scale-in* (instâncias são terminadas).

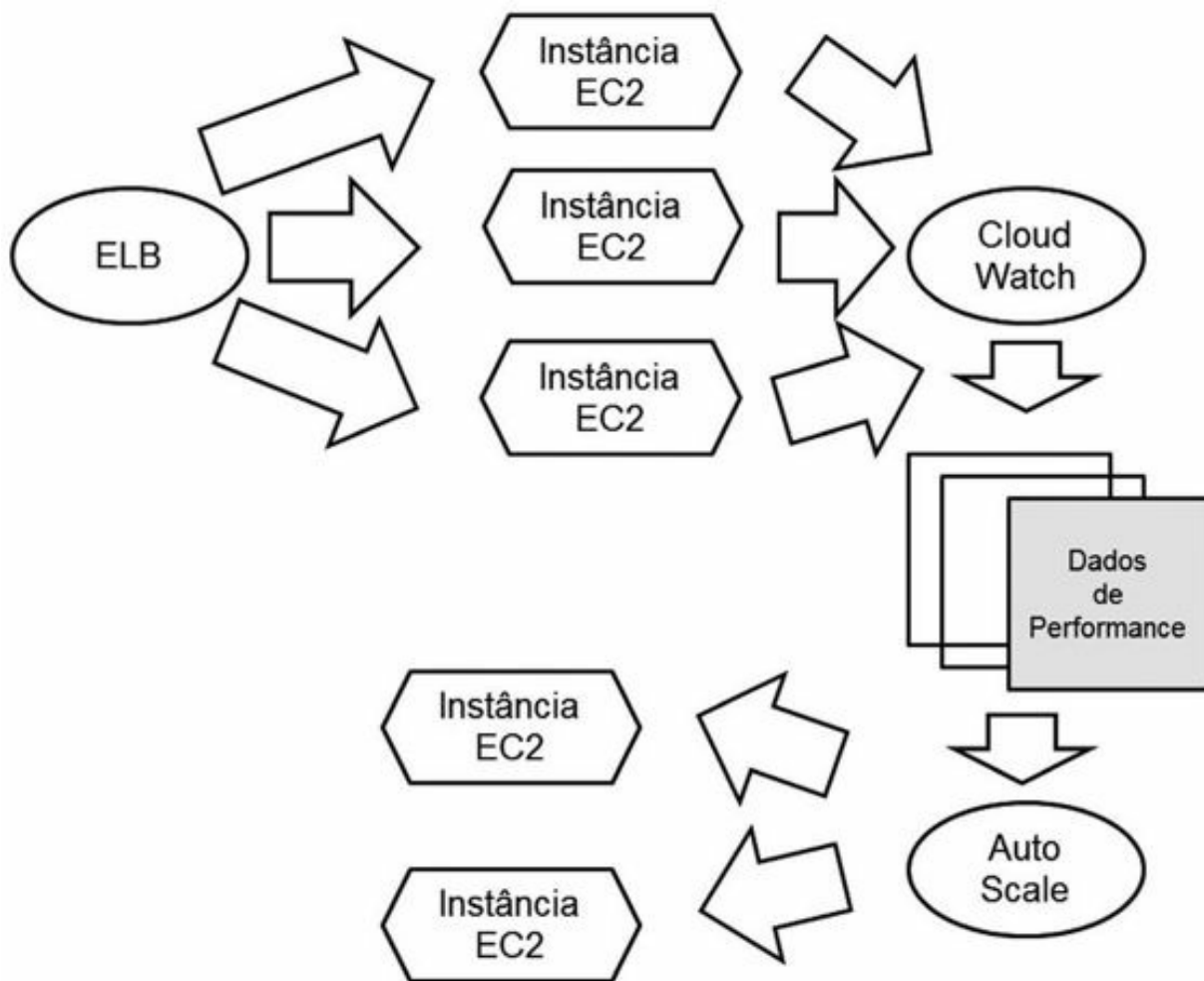


Figura 8-5 – Arquitetura AWS com Auto Scaling

- **Amazon Elastic Load Balancing (Amazon ELB):** o Amazon ELB tráfego de entrada para a aplicação através de várias instâncias EC2. Quando se usa ELB, dá-se um DNS *host name* – qualquer *request* enviado para este host name é delegado para um pool de instâncias EC2. *Auto scaling* e ELB são uma combinação interessante dos serviços Amazon AWS. ELB dá um simples nome DNS para endereçamento e AS estabelece que existe sempre um número certo de instâncias para aceitar as requisições.
- **Amazon Elastic Block Store (Amazon EBS):** o Amazon EBS fornece volumes de armazenamento no nível de bloco para uso com instâncias Amazon EC2. Os volumes Amazon EBS são armazenados e sua persistência é independente da vida de uma instância. O Amazon EBS fornece volumes de armazenamento altamente disponíveis e confiáveis que podem ser anexados a uma instância em execução do Amazon EC2 e expostos como um dispositivo dentro da instância Amazon EBS. Este recurso é particularmente adequado para aplicativos que exigem um banco de dados, sistema de arquivos ou acesso ao armazenamento no nível de bloco bruto. Em resumo, o serviço EBS é uma nova forma de armazenamento persistente criada pela Amazon que permite criar volumes que possam ser anexados como dispositivos em nível de bloco a uma instância em execução. É possível também criar *snapshots* instantâneos a partir desses volumes e depois recriar um volume a partir do *snapshot* instantâneo. Cada *snapshot* instantâneo representa o estado de um volume em um ponto específico no tempo. Dessa

forma, é possível armazenar facilmente arquivos e dados que precisam persistir além do tempo de vida de uma instância em um volume EBS e depois anexar e reanexar esse volume a qualquer instância desejada. A única advertência é que cada volume EBS só pode ser anexado a uma instância por vez. Entretanto, é possível anexar o número desejado de volumes diferentes a uma única instância. Cada volume EBS está localizado em uma zona de disponibilidade, conceito a ser visto posteriormente. A instância a que o volume está sendo anexado deve estar na mesma zona de disponibilidade. Há um limite de conta de 20 volumes EBS, mas é possível solicitar que o Amazon Web Services aumente o limite, se for preciso usar mais volumes. O EBS essencialmente é um volume de disco endereçável. A aplicação pode criar um volume e anexá-lo a uma instância que está rodando na mesma zona de disponibilidade. O volume então pode ser utilizado como se fosse um disco rígido. Vale ressaltar que o volume criado tem o ciclo de vida independente da instância e pode persistir mesmo que a instância deixe de rodar. Pode-se vincular múltiplos volumes EBS a uma instância e pode-se criar um snapshot de um volume. Pode-se criar um novo volume EBS de um *snapshot* e vinculá-lo a outra instância. Pode-se também retirar o vínculo de um volume EBS de uma instância e vinculá-lo a uma instância diferente. A **Figura 8-6** ilustra as opções para o serviço EBS.

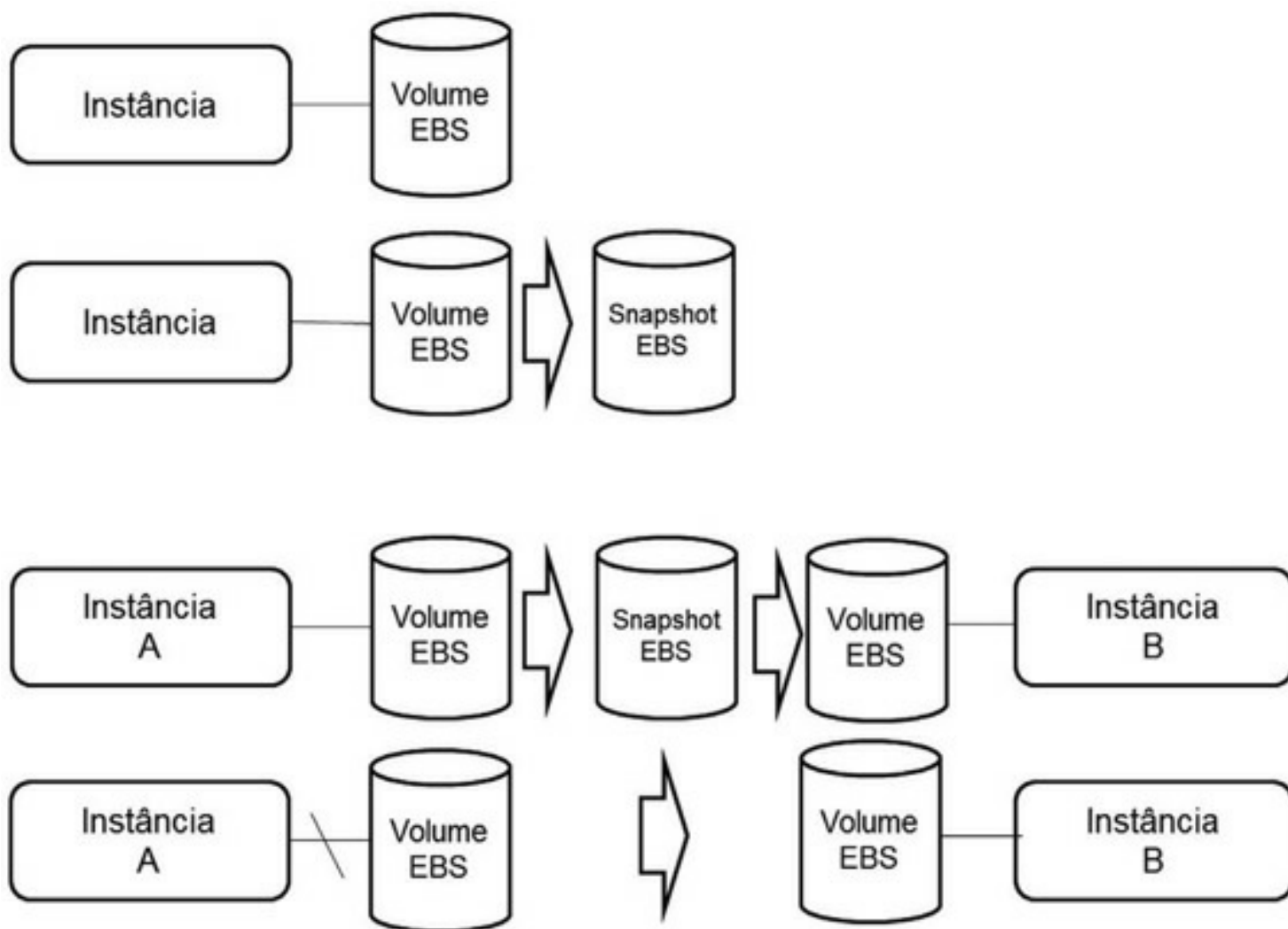


Figura 8-6 – Operações com o Amazon EBS

■ Amazon SimpleDB (Amazon SDB)

O Amazon SDB é um gerenciador de banco de dados do tipo não relacional disponível, escalável e flexível que alivia o trabalho de administração de banco de dados. Os desenvolvedores armazenam itens de dados de consulta via serviços web e o Amazon SimpleDB se encarrega de fazer o gerenciamento, incluindo indexação e consultas.

O Amazon SDB é fácil de usar e fornece a maioria das funções de um banco de dados relacional. A proposta é a de tornar a manutenção mais simples do que a de um banco de dados típico, com a simplificação da configuração. A Amazon cuida de todas as tarefas administrativas. Os dados são automaticamente indexados pela Amazon e estão disponíveis a qualquer hora e em qualquer lugar. Uma vantagem principal de não estar preso a esquemas é a capacidade de inserir dados em tempo real e adicionar novas colunas ou chaves de maneira dinâmica.

Em resumo, o Amazon SimpleDB é otimizado para fornecer alta disponibilidade, facilidade de escalabilidade e flexibilidade, com pouca ou nenhuma carga administrativa. O Amazon SimpleDB pode criar e gerenciar várias réplicas dos dados geograficamente distribuídas para habilitar a alta disponibilidade de dados e a sua confiabilidade. O serviço responde a alterações no tráfego cobrando apenas para recursos de computação e armazenamento efetivamente consumidos. Pode-se alterar o modelo de dados em tempo real e indexar os dados automaticamente. Com o Amazon SimpleDB, pode-se concentrar no desenvolvimento de aplicações sem se preocupar sobre provisionamento de infraestrutura, alta disponibilidade, manutenção de software, gerenciamento de esquema e índice ou ajuste de desempenho.

Domínios

O *domain* é um container que permite armazenar dados estruturados e executar consultas com relação a ele. Os dados são armazenados em um domínio como itens. Conceitualmente, um domínio é similar a uma guia em uma planilha; os itens são linhas na planilha. É possível realizar consultas com relação a um domínio, mas ainda não é possível fazer consultas entre domínios na versão atual do SDB.

O SDB segue o modelo “consistência eventual”. O SDB mantém várias cópias de cada domínio para tolerância a falhas. Toda mudança feita em um domínio é propagada em todas as cópias.

Como essa operação às vezes leva alguns segundos, dependendo da carga do sistema e da latência de rede, um consumidor do seu domínio pode não ver as alterações imediatamente. As mudanças serão finalmente propagadas em todo o SDB, mas esse atraso é uma importante consideração de quando projetar os aplicativos baseados em SDB.

Itens

A versão atual do SDB tem limitações que devem ser levadas em consideração no projeto do aplicativo. A **Tabela 8-1** mostra as limitações (como especificado pelo Amazon em sua documentação mais recente).

Tabela 8-1 – Limitações atuais do SDB

Parâmetro	Restrições atuais
Tamanho do domínio	10 GB por domínio 250.000.000 pares de nome-valor de atributo 3-255 caracteres (a-z, A-Z, 0-9, ‘_’, ‘-’ e ‘.’)
Domínios por conta Amazon Web Services	100
Atributos	Pares nome-valor por item são 256. Comprimento do nome é de 1024 bytes. Comprimento do valor é de 1024 bytes. Somente os caracteres permitidos são os UTF-8 válidos em documentos XML. Caracteres de controle e todas as sequências não válidas no XML não são permitidos. Por operação PutAttributes limitada a 100 Solicitado por operação Select ou QueryWithAttributes limitada a 256.
Máximo de itens na resposta de consulta	256
Tempo máximo de execução da consulta	5 segundos
Máximo de predicados por expressão de consulta	10
Máximo de comparações por predicado de expressão de consulta	10
Número máximo de atributos exclusivos por expressão de seleção	20
Número máximo de comparações por expressão de seleção	20
Tamanho máximo da resposta para QueryWithAttributes e Select	1 MB

■ Amazon Simple Storage Service (Amazon S3)

O Amazon S3 é um armazenamento de dados para a Internet. Ele foi projetado para fazer a utilização da web mais fácil para desenvolvedores. O Amazon S3 fornece uma interface de serviços web simples que pode ser usada para armazenar e recuperar qualquer quantidade de dados, a qualquer momento, de qualquer lugar na web, segundo a Amazon. O S3 permite que o desenvolvedor utilize uma infraestrutura escalável, confiável, segura, rápida e barata que a Amazon mesmo usa para executar sua própria rede global de sites da web. O serviço visa maximizar os benefícios da escala e passar esses benefícios para os desenvolvedores.

O Amazon S3 é um sistema de armazenamento de dados na Internet escalável e rápido que simplifica o armazenamento e recuperação de qualquer volume de dados, a qualquer momento e de qualquer parte do mundo.

No Amazon S3 paga-se pelo armazenamento e pela largura de banda com base no uso real do serviço. Não há custo de instalação, custo mínimo ou custo recorrente adicional.

A Amazon fornece a administração e a manutenção da infraestrutura de armazenamento, deixando o desenvolvedor livre para se concentrar nas funções principais de seus sistemas e aplicativos. S3 é uma plataforma abrangente que está disponível para necessidades de armazenamento de dados. Ela é excelente para:

- Armazenar dados dos aplicativos.
- Backups.
- Distribuição rápida e barata de mídias e outros conteúdos que consomem muita largura de banda.

Os três conceitos básicos que sustentam a estrutura de trabalho da S3 são *buckets*, objetos e chaves.

Buckets

Buckets são CONTAINERS para objetos S3. Cada objeto armazenado no S3 está contido em um *bucket*. O *bucket* é uma espécie de diretório no sistema de arquivos. Uma das principais distinções entre uma pasta de arquivos e um *bucket* é que cada *bucket* e seu conteúdo podem ser acessados usando uma URL.

Cada conta da S3 pode conter o máximo de 100 *buckets*. *Buckets* não podem ser aninhados um dentro do outro, então não é possível criar um *bucket* dentro de outro *bucket*.

É possível afetar a localização geográfica dos *buckets* especificando uma restrição de localização ao criá-los. Isso automaticamente garantirá que quaisquer objetos armazenados dentro daquele *bucket* sejam armazenados naquela localização geográfica. Se não se especificar uma localização ao criar o *bucket*, o *bucket* e seu conteúdo serão armazenados na localização mais próxima do endereço de cobrança da conta.

Nomes de bucket precisam atender aos seguintes requisitos do S3:

- O nome deve iniciar com um número ou uma letra.
- O nome deve ter entre 3 e 255 caracteres.
- Um nome válido pode conter somente letras minúsculas, números, pontos, sublinhados e travessões.
- Apesar de ser possível usar números e pontos no nome, eles não podem ter o formato de um endereço IP. Não é possível dar o nome 192.168.1.254 a um *bucket*.
- O espaço de nome do *bucket* é compartilhado com todos os *buckets* de todas as contas na S3. O nome deve ser exclusivo em toda a S3.

Objetos

Objetos contêm os dados armazenados dentro de *buckets* na S3. Pense em um objeto como o arquivo que se deseja armazenar. Cada objeto armazenado é composto de duas entidades: dados e metadados.

Os dados são a coisa real a ser armazenada, como um arquivo PDF, um documento do Word, etc. Os dados armazenados também têm metadados associados para descrever o objeto.

Os metadados são dados sobre dados. Armazenam o tipo de conteúdo do objeto sendo armazenado, a data em que o objeto foi modificado pela última vez e quaisquer outros metadados específicos do usuário ou do aplicativo. Os metadados de um objeto são especificados pelo desenvolvedor como pares chave-valor quando o objeto é enviado ao S3 para armazenamento.

Diferentemente da limitação no número de *buckets*, não existem restrições no número de objetos. É possível armazenar um número ilimitado de objetos nos *buckets* e cada objeto pode conter até 5 GB de dados. Os dados nos objetos S3 publicamente acessíveis podem ser recuperados por HTTP, HTTPS ou BitTorrent.

Chaves

Cada objeto armazenado dentro de um *bucket* da S3 é identificado usando uma chave exclusiva. Isso é similar em conceito ao nome de um arquivo em uma pasta de um sistema de arquivos.

O nome do arquivo dentro de uma pasta em um disco rígido deve ser exclusivo. Cada objeto dentro de um *bucket* tem uma chave exclusiva. O nome do *bucket* e da chave são usados em conjunto para fornecer a identificação exclusiva de cada objeto armazenado na S3.

O endereço de cada objeto dentro da S3 é uma URL que combina a URL de serviço da S3, o nome do *bucket* e a chave exclusiva. Apesar de serem conceitos simples, *buckets*, objetos e chaves, juntos, fornecem muita flexibilidade para criar soluções de armazenamento de dados.

É possível aproveitar esses blocos modulares para simplificar o armazenamento de dados na S3 ou usar da flexibilidade para criar camadas e construir armazenamento ou aplicativos mais complexos sobre a S3 para fornecer funções adicionais.

O S3 pode armazenar objetos de 5 TB de tamanho e tem um modelo de precificação específico que depende de quantidade de dados armazenados, da quantidade de dados transferidos para e do S3 e do número de requests feitos ao S3.

■ Amazon Simple Queue Service (Amazon SQS)

A mensageria assíncrona é o elemento chave para a construção de aplicações escaláveis. Com a mensageria pode-se construir as aplicações de forma modular e um *pipeline* controla as mensagens a serem enviadas entre os passos de processamento da aplicação. O Amazon SQS é o serviço de mensageria do Amazon EC2.

O Amazon SQS fornece acesso à infraestrutura do sistema de mensagens usada pela Amazon. É possível enviar e recuperar mensagens de qualquer lugar usando pedidos HTTP baseados em REST. Não é necessário instalar nem configurar nada.

Com o SQS é possível criar um número ilimitado de filas e enviar um número ilimitado de mensagens. As mensagens são armazenadas pela Amazon em servidores e DATACENTERS múltiplos para fornecer a redundância e a confiabilidade necessária para um sistema de mensagens.

O SQS integra-se aos outros serviços web da Amazon. Ele fornece uma maneira de criar um sistema desacoplado onde as instâncias EC2 podem comunicar-se umas com as outras enviando

mensagens ao SQS e permite coordenar o fluxo de trabalho.

Também é possível usar as filas de mensagem para construir uma infraestrutura com escalabilidade e recuperação automática baseada em EC2 para o aplicativo. É possível tornar as mensagens seguras em uma fila contra o acesso não autorizado usando os mecanismos de autenticação fornecidos pelo SQS.

É possível usá-lo como base para unir os aplicativos baseados no Amazon Web Services. Usar SQS é uma excelente forma de criar aplicativos web realmente desacoplados. Pagam-se pelas mensagens totalmente com base no uso. Toda a estrutura de enfileiramento é executada dentro do ambiente seguro dos DATACENTERS próprios da Amazon.

Mensagens

Mensagens contêm dados de texto de até 8 KB. Cada mensagem é armazenada até ser recuperada pelo aplicativo de recebimento. Um valor de tempo limite de visibilidade, em segundos, é especificado quando o aplicativo de recebimento lê uma mensagem de uma fila. Isso funciona mais como uma trava e assegura que, pelo período especificado:

- A mensagem recuperada não estará disponível a qualquer outro consumidor da fila.
- A mensagem somente reaparecerá na fila quando o período de tempo limite expirar se, e somente se, ela não tiver sido excluída pelo processo de leitura.
- As mensagens ficam retidas em uma fila por quatro dias. O SQS excluirá automaticamente todas as mensagens que estiverem nas filas por mais de quatro dias.

O SQS segue o modelo de “consistência eventual”, significando que é possível enviar uma mensagem à fila, mas um consumidor dessa fila pode não ver a mensagem por um período significativo. A mensagem será finalmente entregue, mas essa será uma consideração importante se o seu aplicativo levar em conta a ordem das mensagens.

Filas

Filas são CONTAINERS para as mensagens. Cada mensagem deve especificar uma fila que a manterá. As mensagens enviadas a uma fila permanecem nela até elas serem explicitamente excluídas. A ordem da fila é a primeira a entrar e sair, mas a ordem não é garantida.

Cada fila tem um tempo limite de visibilidade padrão de 30 segundos. É possível alterar esse valor para toda a fila ou ele pode ser definido individualmente para cada mensagem sendo recuperada.

O tempo limite máximo de visibilidade de uma fila ou mensagem é de duas horas (7200 segundos). O SQS reserva-se no direito de excluir automaticamente as filas, caso não haja nenhuma atividade na fila por 30 dias corridos.

Design

O SQS é um pouco diferente das estruturas de fila comuns. Há três coisas que devem ser consideradas antes de projetar os aplicativos baseados em SQL:

- O SQS não garante a ordem das mensagens em uma fila.
- As mensagens são colocadas de maneira desordenada na fila. Elas não são realmente armazenadas na ordem em que são adicionadas à fila. O SQS tentará preservar a ordem das mensagens, mas não há garantia de que você receberá mensagens na ordem exata em que elas são enviadas. Se a ordem das mensagens for importante para o seu aplicativo, será preciso adicionar dados de sequência a cada mensagem.
- O SQS não garante a exclusão de uma mensagem na fila.

O aplicativo deve ser projetado de forma que ele não seja afetado se a mesma mensagem for processada mais de uma vez. Todas as mensagens são armazenadas em vários servidores pelo SQS para fornecer redundância e alta disponibilidade. Se um desses vários servidores ficar indisponível enquanto uma mensagem estiver sendo excluída, é possível, em raras circunstâncias, obter uma cópia da mensagem novamente enquanto estiver recuperando mensagens.

O SQS não garante que todas as mensagens da fila serão retornadas quando consultadas.

O SQS usa amostragem de mensagem com base em distribuição aleatória ponderada e retorna mensagens somente do subconjunto amostrado de servidores quando você consulta as mensagens. Embora uma determinada solicitação possa não retornar todas as mensagens da fila, se continuar recuperando mensagens da fila, ocorrerá a amostragem de todos os servidores e serão recebidas todas as mensagens.

■ **Outros serviços Amazon AWS:**

- **Amazon Relational Database Service (Amazon RDS):** O Amazon RDS torna mais fácil de configurar, operar e dimensionar um banco de dados relacional na nuvem. O serviço fornece capacidade a um custo adequado e é redimensionável enquanto gerencia tarefas de administração de banco de dados normalmente demoradas, liberando o analista para se concentrar em seus aplicativos e negócios. O Amazon RDS permite acesso a todos os recursos do banco de dados MySQL. Isso significa que código, aplicações e ferramentas que o usuário já usa hoje em bancos de dados MySQL podem continuar existindo no Amazon RDS. O Amazon RDS aplica automaticamente patches do software de banco de dados e faz o backup do banco de dados, armazenando-os por um período de retenção definido pelo usuário. No RDS pode-se beneficiar da flexibilidade de poder dimensionar os recursos de computação ou capacidade de armazenamento associado com a instância do banco de dados relacional através de uma única chamada de API. Além disso, o Amazon RDS permite implantar a instância de banco de dados em várias zonas de disponibilidade para conseguir melhorar a disponibilidade e confiabilidade para implantações de produção críticos.
- **Amazon CloudFront:** é um serviço web para disponibilização de conteúdo. Integra-se com outros Amazon Web Services e dá aos desenvolvedores um caminho fácil para distribuir conteúdo para usuários finais com baixa latência e transferência de dados em alta velocidade. Os *requests* dos objetos são automaticamente roteados para a localização mais próxima, otimizando a performance. O CloudFront hoje existe em dezenove localizações e foi projetado para trabalhar junto com o Amazon S3.

Basicamente o CloudFront reduz a latência de acessar os dados diretamente no *bucket* do S3. Os acessos podem ser feitos via HTTP, HTTPS ou RTMP. O CloudFront tem sua precificação baseada na quantidade de dados transferidos e no número de *requests* feito ao CloudFront.

- **Amazon Virtual Private Cloud (VPC):** é uma ponte segura entre a infraestrutura de TI de uma organização e a Nuvem AWS. A Amazon VPC permite conectar, através de uma conexão VPN, a infraestrutura interna com recursos de computação AWS isolados. Sistemas de firewall, sistemas de detecção de intrusão passam a incluir os recursos AWS. O Amazon VPS se integra com Amazon EC2. A **Figura 8-7** ilustra a utilização do serviço Amazon VPC para uma empresa com DATACENTER próprio.

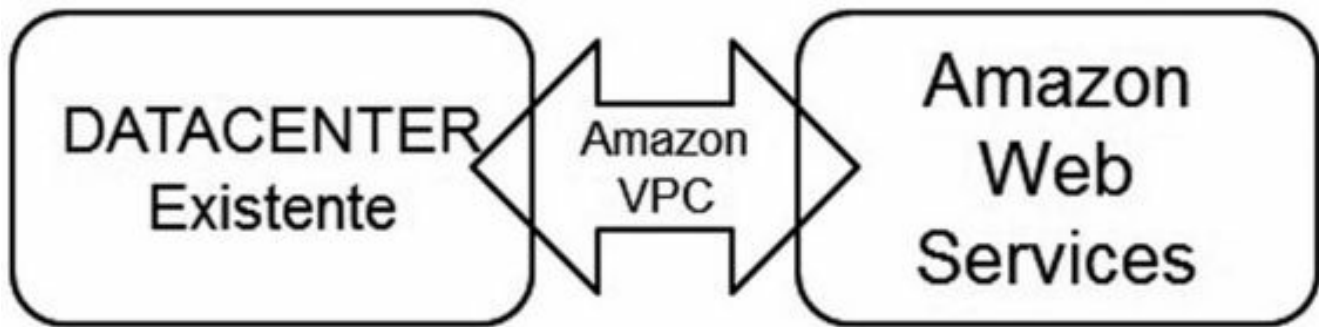


Figura 8-7 – Amazon VPC

- **Amazon Route 53:** é um serviço DNS altamente disponível e escalável que dá aos desenvolvedores um caminho para rotear usuários finais para aplicações Internet fazendo a tradução de nomes para endereços IP.
- **Amazon Simple Notification Service (SNS):** facilita o envio de notificações da nuvem. Desenvolvedores podem publicar mensagens de uma aplicação e imediatamente disponibilizá-las para assinantes ou outras aplicações.
- **Amazon Elastic MapReduce:** possibilita o processamento de grandes quantidades de dados. Utiliza um *framework* Hadoop rodando na infraestrutura do EC2 e Amazon S3.
- **Amazon ElastiCache:** é um serviço web que permite a utilização de dados em memória cache e não em disco por aplicações web utilizadas com sistemas gerenciadores de banco de dados. ElastiCache é *compliance* com o protocolo Memcached.
- **Amazon AWS Import/Export:** acelera o movimento de grandes quantidades de dados dentro e fora do AWS utilizando *buckets* do S3. Para isto utiliza a rede interna de alta performance e *bypassa* a Internet. O serviço só existe para algumas regiões.
- **Amazon Toolkit for Eclipse:** é um plug-in para Eclipse (Java IDE) que torna fácil a vida de desenvolvedores Java, ajudando a desenvolver, debugar e pôr em produção aplicações Java que usam o AWS.
- **Amazon AWS Elastic Beanstalk:** é um caminho mais fácil para desenvolver e pôr em

produção aplicações na nuvem AWS. Quando se faz o upload da aplicação, o Elastic Beanstalk automaticamente fornece o provisionamento da capacidade, o balanceamento de carga, o *auto-scaling* e o monitoramento para a aplicação.

- **O AWS Architecture Center:** fornece informações sobre como projetar a arquitetura de aplicações para o AWS com foco em escalabilidade e confiabilidade. Maiores informações podem ser encontradas em <http://aws.amazon.com/architecture/>.

8.2.3. Centros de Suporte dos Serviços AWS

A Amazon possui dois centros com foco em aspectos de custo e segurança. O foco destes centros é o de ajudar os clientes a entenderem as reais vantagens do ambiente de nuvem em termos de custo, benefício e segurança.

AWS Economics Center

O AWS Economics Center fornece acesso a informações, ferramentas e recursos para comparar os custos da Amazon Web Services com alternativas convencionais de infraestrutura de TI. O objetivo é ajudar os desenvolvedores e líderes empresariais a quantificar os benefícios econômicos da CLOUD COMPUTING.

A Amazon desenvolveu o Amazon EC2 Cost Comparison Calculator, uma planilha Excel que ajuda os decisores a quantificarem os benefícios da CLOUD COMPUTING quando comparado à tradicional infraestrutura alternativa.

A última versão da planilha pode ser obtida em: <http://aws.amazon.com/economics>

O artigo The Economics of the AWS CLOUD vs. Owned IT Infrastructure, publicado em 2009 e disponível na web, compara os custos diretos e indiretos para as duas opções.

AWS Security Center

O AWS Security Center fornece links para informações técnicas, ferramentas e orientação para ajudar a criar e gerenciar aplicativos seguros na nuvem AWS. O objetivo é usar este fórum para notificar proativamente os desenvolvedores sobre boletins de segurança. O site da Amazon com informações de segurança é este:

<http://aws.amazon.com/security>

AWS também faz backup dos dados armazenados no Amazon S3, Amazon Simple-DB ou Amazon Elastic Block Store (EBS). Os dados são armazenados de forma redundante em múltiplas localizações físicas como parte da operação normal.

8.2.4. Funcionamento do Amazon AWS EC2

Para entender o funcionamento do Amazon AWS é necessário conhecer alguns conceitos descritos a seguir:

- **Recursos Persistentes e Efêmeros.** Os recursos do Amazon EC2 podem ser agrupados em duas classes: persistentes e efêmeros.
 - Recursos persistentes permanecem operacionais mesmo que aconteça uma falha de

hardware ou uma falha de software. São eles: endereços IP elásticos, volumes EBS, ELBs, grupos de segurança e AMIs armazenadas no S3 ou como EBS snapshots.

- Recursos efêmeros não possuem redundância e podem falhar. Quando falham as informações armazenadas, o estado da informação é geralmente perdido. Assim, deve-se garantir a disponibilidade destes recursos utilizando outras funcionalidades. O Amazon EC2 é efêmero. Deve-se pensar esta característica para o EC2 como uma funcionalidade e não como bug pois o AWS é todo baseado nesta condição.
- **Pares de Chave de Segurança.** São pares de chave SSH públicos/privados especificados na ativação de uma instância. Eles são necessários para na verdade fazer login no console de uma de suas instâncias ativadas. Os pares de chave de segurança são utilizados principalmente para permitir que os usuários façam login com segurança em instâncias sem a necessidade do uso de senhas. Os pares de chave de segurança são diferentes do ID da chave de acesso e da chave de segurança do Amazon Web Services (disponíveis na página de dados da conta), que são usados para identificar você exclusivamente como o usuário que faz as solicitações ao Amazon Web Services usando a API.
- **Grupo de Segurança.** Toda e qualquer instância iniciada no ambiente EC2 é executada dentro de um grupo de segurança. Cada grupo de segurança define as regras de firewall que especificam as restrições de acesso para instâncias executadas dentro desse grupo. É possível conceder ou restringir o acesso por endereço IP ou regras *Classless Interdomain Routing* (CIDR), que permitem especificar um intervalo de portas e o protocolo de transporte. Também é possível controlar o acesso a grupos de segurança especificados, para que qualquer instância em execução nesses grupos de acesso de segurança tenha acesso automaticamente concedido ou negado à instância.
- **Amazon Machine Image (AMI).** É uma imagem (*template*) que contém uma configuração de sistema operacional e possivelmente outros softwares como gerenciadores de banco de dados, *middleware* e servidores web. É similar ao drive raiz de um computador. O serviço EC2 permite lançar instâncias que são cópias da AMI. A Amazon publica muitas AMIs contendo configurações comuns de software para uso público. Membros da comunidade de desenvolvedores também publicam suas próprias AMIs. Existem três tipos de AMIs, ilustradas na **Tabela 8-2**.

Tabela 8-2 – Tipos de AMIs

Tipo	Definição
Privada	Imagens criadas por você, que são particulares por padrão. É possível conceder acesso a outros usuários para iniciar suas imagens privadas.
Pública	Imagens criadas pelos usuários e liberadas para a comunidade do Amazon Web Services, para que qualquer pessoa possa iniciar instâncias com base

nelas e usá-las da forma que desejar. O Amazon Web Services Developer Connection lista todas as imagens públicas.

Paga	É possível criar imagens fornecendo funções específicas que podem ser iniciadas por qualquer pessoa que queira pagá-las por hora de uso com base nas taxas do serviço.
------	--

A Amazon AWS fornece a opção de linha de comando, que facilita, a criação e o gerenciamento de AMIs. As próprias AMIs são armazenadas no Amazon Simple Storage Service (S3).

Após o registro da imagem com o EC2, um ID exclusivo é atribuído à imagem, que pode ser usado para identificá-la e iniciar uma instância a partir dela. Há várias formas de criar sua própria imagem.

Pode-se lançar múltiplas instâncias da AMI. De uma AMI pode-se também lançar diferentes tipos de instâncias. A [Figura 8-8](#) ilustra estas possibilidades com a AMI.

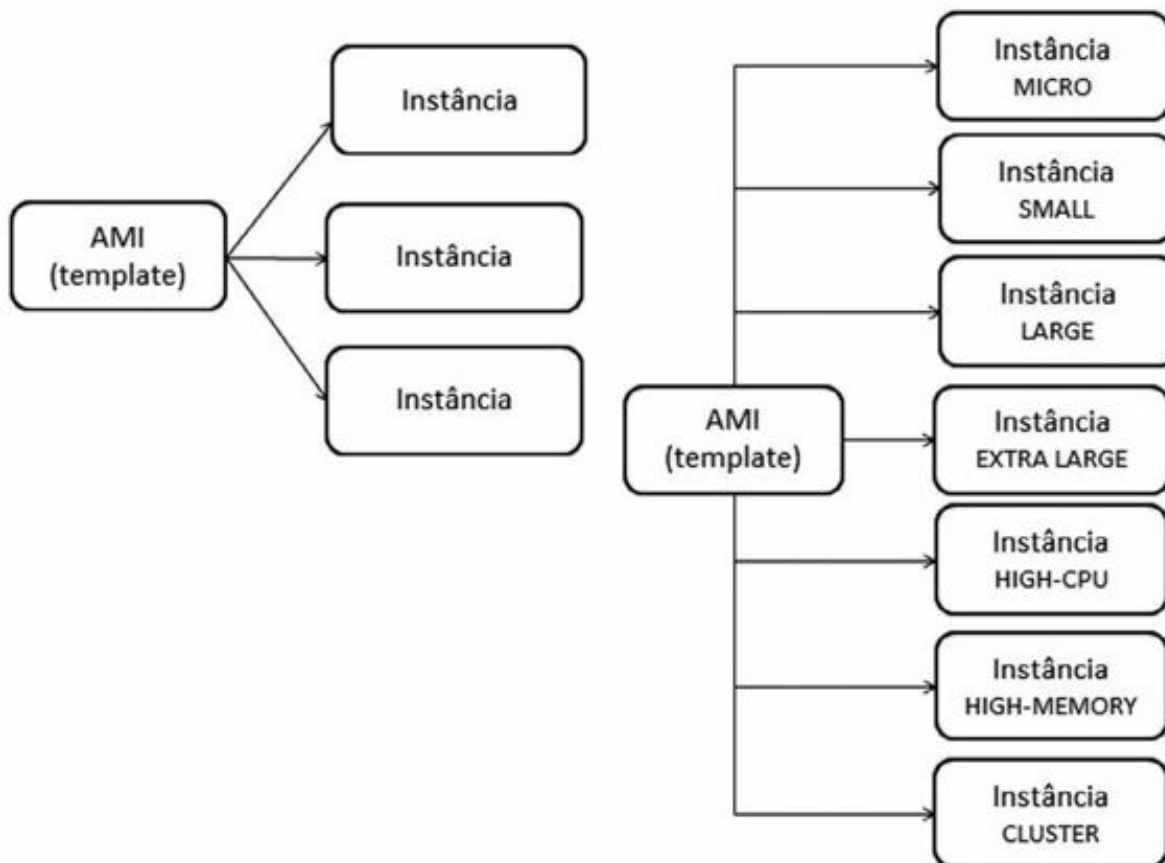


Figura 8-8 – Forma de instanciar AMIs.

Instâncias

São as instâncias virtuais em execução que usam uma AMI como modelo. É possível iniciar uma instância, visualizar detalhes sobre ela e encerrá-la usando as ferramentas fornecidas pela Amazon. É possível também usar uma variedade de bibliotecas de terceiros em diferentes linguagens

para controlar o seu ciclo de vida.

O EC2 roda em cima de uma plataforma de virtualização baseada no hypervisor XEN. Os sistemas operacionais convidados (*guests*) que compõem as instâncias podem ser: Windows Server, RedHat Enterprise Linux, Open Solaris, Suse Linux Enterprise, Ubuntu Linux, Debian, Oracle Enterprise Linux, Fedora, Gentoo Linux e Amazon Linux AMI.

Como as instâncias são cobradas com base no tempo real de uso, é possível dimensionar facilmente as necessidades de computação com base na carga atual do aplicativo. Não é preciso obter a capacidade antecipadamente.

Opções de compra de instâncias

A Amazon AWS possibilita adquirir as instâncias de quatro principais formas, permitindo que o cliente otimize os custos e tenha flexibilidade.

- **Instâncias *on-demand*.** São instância adquiridas sob demanda do AWS e pagas por hora. Não existe compromisso antecipado em adquirir as instâncias nem custos extras. Instâncias de aplicações em geral são adquiridas desta forma.
- **Instâncias reservadas.** Permitem reservar capacidade computacional no AWS considerando que os recursos são escassos. Essas reservas são feitas baseadas em contratos de um a três anos e garantem a disponibilidade de instâncias em regiões e zonas escolhidas. Esta funcionalidade reduz o preço e permite conseguir capacidade computacional quando a organização precisar. Aplicações que utilizam *steady-state* são potencialmente interessantes para esta modalidade de aquisição.
- **Instâncias *spot*.** Podem ser adquiridas a qualquer momento por um período de tempo e possibilitam definir o preço por hora para um tipo de instância e para rodar em uma zona de disponibilidade específica. O preço flutua com base na oferta e na demanda.
- **Instâncias dedicadas.** Rodam em máquinas exclusivas por questões de segurança ou *compliance*.

Endereços IP Elásticos

A cada instância é atribuída automaticamente um endereço IP privado e um público na ativação realizada pelo EC2. O endereço IP público pode ser usado para acessar a instância pela Internet. Entretanto, sempre que for ativada uma instância, esse endereço mudará.

Para o uso de qualquer tipo de mapeamento de DNS dinâmico para conectar um nome DNS ao endereço IP, a alteração poderá levar até 24 horas para ser propagada pela Internet. O EC2 introduziu o conceito de endereço IP elástico para aliviar esse problema.

Endereços IP elásticos são IPs públicos que podem ser mapeados para qualquer instância EC2 e dentro uma região EC2. Os endereços são associados com uma conta AWS e não com uma instância específica ou o tempo de vida de uma aplicação e é ideal para construir aplicações tolerantes a falhas. A operação elástica dos IPs é realizada através de uma API ou pelo console.

Cada endereço IP elástico é associado a sua conta do EC2 e não a uma instância específica e será associado permanentemente à conta, a menos que você explicitamente o libere novamente para

EC2.

É possível também remapear um endereço IP elástico entre instâncias e dessa forma responder rapidamente a qualquer falha de instância apenas iniciando outra instância e remapeando-a (ou usando uma instância existente). A qualquer momento é possível apenas ter uma única instância mapeada para um endereço IP elástico.

AWS Management Console

O AWS Management Console é parte do AWS. Quando logado, o usuário tem acesso de forma simples aos principais serviços AWS, incluindo o EC2 e o S3. Pode-se gerenciar grupos de segurança, endereços IP elásticos, volumes EBS, lançar novas instâncias e ver as instâncias que estão rodando, além de identificar pares de chave.

A forma mais fácil de interagir com o AWS é através do AWS Management Console. Existe também a opção por linha de comando ou mesmo através de APIs, mas com o Management Console tudo fica mais fácil.

A **Figura 8-9** ilustra a tela do AWS Management Console.

EC2

EC3

VPC

CloudWatch

Elastic MapReduce

CloudFront

CloudFormation

IoT

IoT Greengrass

IoT Device Defender

IoT Fleet Hub

IoT SiteWise

IoT Things

IoT TwinMaker

IoT Analytics

IoT RoboMaker

IoT Device Defender

IoT Fleet Hub

IoT SiteWise

IoT Things

IoT TwinMaker

IoT Analytics

IoT RoboMaker

Region: us-east-1 (Virginia)

INSTANCES

Spot Requests

Reserved Instances

IMAGES

AMIs

Bundle Tasks

ELASTIC BLOCK STORE

Volumes

Environments

Launch Instance

Instance Actions

My Instances

Download

Refresh

Help

1 to 2 of 2 instances

Navigation

INSTANCES

Spot Requests

Reserved Instances

IMAGES

AMIs

Bundle Tasks

ELASTIC BLOCK STORE

Volumes

Name	Instance	AMI ID	Root Device	Type	Status	Security Groups	Key Pair Name	Monitoring	Virtualization	Placement
☐ empty	i-d6ba8b6	ami-0c1f3ce5	ebs	t1.micro	running	quick-start-2	manuel	basic	paravirtual	
☐ empty	i-24756d44	ami-0c1f3ce5	ebs	m1.small	running	quick-start-2	manuel	basic	paravirtual	

0 EC2 instances selected

Select an instance above

Uma sequência típica para um usuário iniciante no AWS é lançar uma instância baseada em uma AMI pública, testá-la e depois desligá-la. A sequência para uma operação deste tipo é mostrada na **Figura 8-10**.

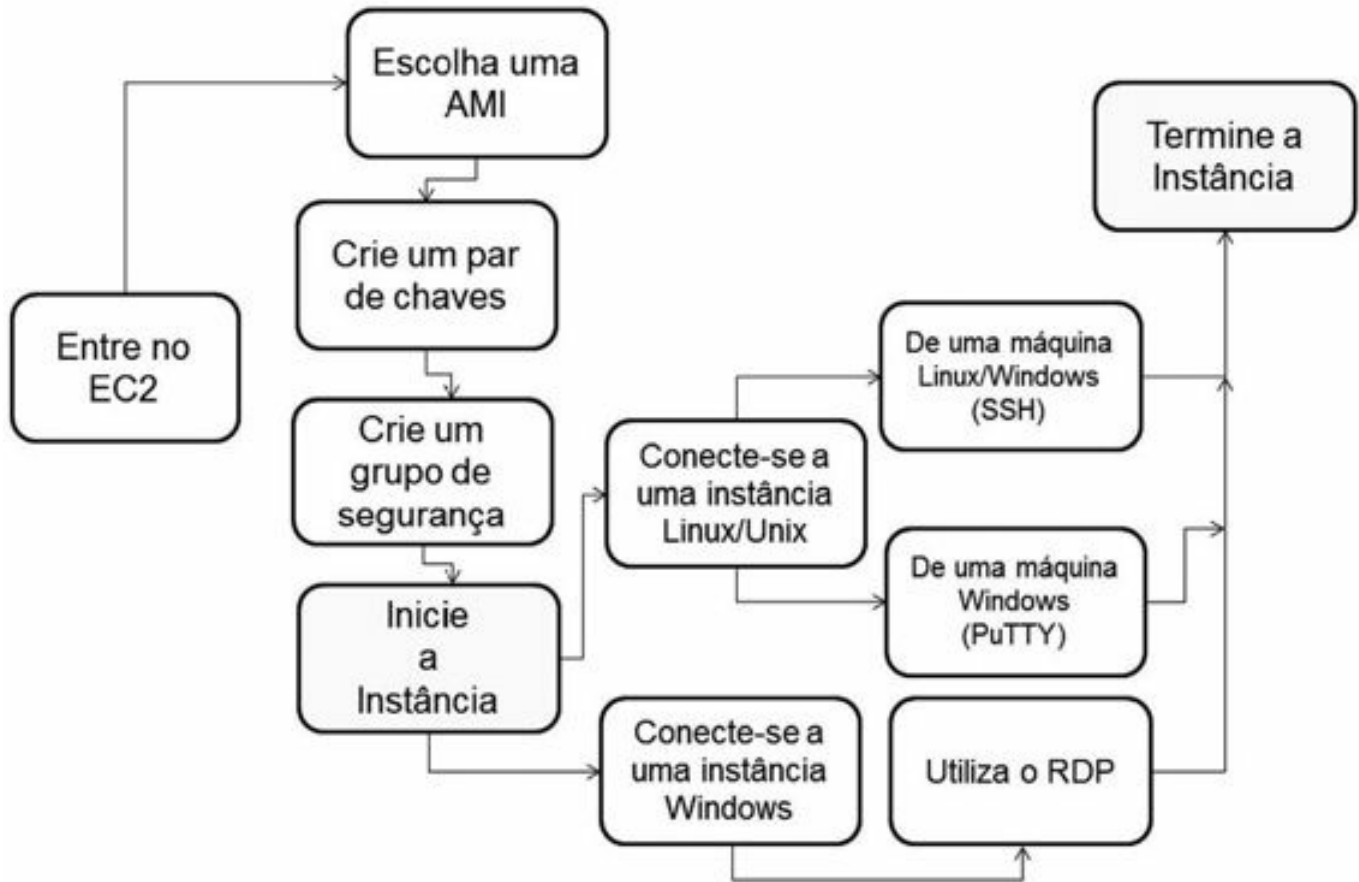


Figura 8-10 – Utilizando o AWS

Observe que todos os passos podem ser dados utilizando o console. A forma de acessar a instância depende do tipo de sistema operacional da instância e do tipo de sistema operacional da máquina cliente.

8.2.5. Aplicações Tolerantes a Falhas no AWS

Introdução

Tolerância a falhas é a habilidade de um sistema de permanecer em operação mesmo se alguns dos componentes usados para construir o sistema falham.

A proposta da plataforma AWS, diferentemente de uma operação convencional de DATACENTER, é que seja possível construir aplicações tolerantes a falhas que operem com uma interação mínima humana, maximizando o investimento.

O serviço fundamental para a construção de aplicações tolerantes a falhas é o Amazon EC2,

que responde no AWS pelos recursos de computação – literalmente instâncias de servidores – utilizados para construir e hospedar as aplicações. O Amazon EC2 é o ponto de entrada para o desenvolvimento de uma aplicação. EC2 é similar ao tradicional servidor. Pode-se construir aplicações tolerantes a falhas usando múltiplas instâncias do EC2.

As instâncias do EC2 utilizam SOs conhecidos como Windows e Linux; assim podem acomodar de forma muito próxima softwares construídos para estes SOs.

Outro serviço fundamental para a alta disponibilidade é o Amazon EBS, onde são armazenadas instâncias *off-line* que persistem independentemente da vida da instância. EBS são, na prática, HDs que podem ser attachados a uma instância Amazon EC2 que está rodando. O EBS armazena dados de forma redundante, reduzindo a falha anual em comparação com um HD convencional (0,5% contra 4%).

O Amazon EBS é confiável, mas mesmo assim é possível criar backup dos seus volumes através da técnica de snapshot. Alguns cuidados devem ser tomados com os dados em memória antes do início do snapshot. O Amazon EBS pode ser anexado a outra instância EC2, caso a primeira instância falhe.

Este ano, com a queda de um dos DATACENTERS da Amazon e a consequente parada dos serviços, ficou evidente que as aplicações precisam ser construídas visando obter alta disponibilidade. O link <http://www.awsdowntime.com/> mostra o downtime anual do serviço AWS. Regiões e zonas de alta disponibilidade são o segredo para a construção de aplicações altamente disponíveis.

Regiões e zonas de disponibilidade

AWS EC2 são disponíveis atualmente em seis regiões. Quando se utiliza o AWS pode-se definir onde os dados devem ficar armazenados, onde rodam as instâncias, onde as filas são inicializadas e banco de dados instanciados. Dentro de cada região existem zona de disponibilidade. Zonas de disponibilidade são locais que permitem que as aplicações sejam altamente disponíveis dentro de uma região. Regiões consistem de uma ou mais zonas de disponibilidade dispersas que são separadas em áreas geográficas diferentes ou países. O SLA para cada região é de 99.95%. Os endereços elásticos IP são críticos no projeto de uma aplicação tolerante a falhas baseado em zonas de disponibilidade separadas. Instâncias redundantes para cada camada de aplicação devem ser colocadas em zonas diferentes.

Instâncias Amazon EC2 podem rodar em múltiplas zonas de disponibilidade. Pode-se utilizar ELB para conseguir alta disponibilidade com baixa intervenção. O ELB pode balancear tráfego entre múltiplas instâncias e múltiplas zonas de disponibilidade.

O EC2 é composto por vários DATACENTERS em localidades separadas para fornecer resiliência a falhas. É possível colocar as instâncias ativadas em locais diferentes. Os locais são regiões geográficas com zonas de disponibilidade dentro delas.

É possível proteger aplicativos contra a falha de um único local ativando instâncias em zonas de disponibilidade separadas. Se não se especificar uma zona de disponibilidade ao iniciar uma instância, o Amazon EC2 escolherá automaticamente uma para você com base no funcionamento do sistema e na capacidade atuais.

Outros serviços como SQS, S3, SimpleDB, RDS são blocos de construção de alto nível que

podem ser incorporados à aplicação e melhorar ainda mais o nível de disponibilidade. A **Figura 8-11** mostra o local das Regiões e Zonas de Disponibilidade.

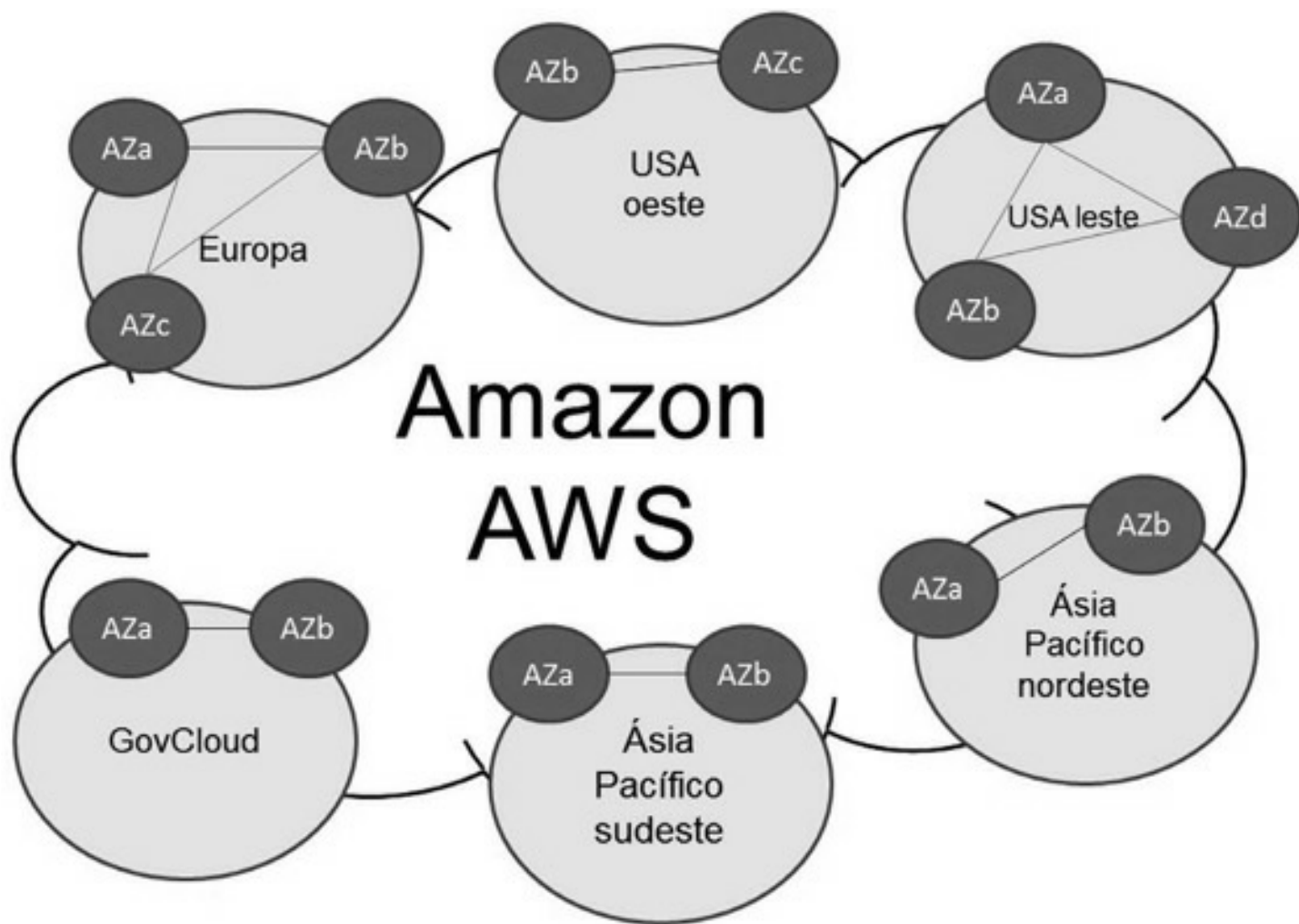


Figura 8-11 – Regiões e Zonas de Disponibilidade

A região AWS GovCloud obedece aos seguintes requisitos:

- Usuários devem ser cidadãos americanos.
- Não estar sujeito a restrições de exportações.
- Estar *compliance* com leis de controle e regulações de exportação.

8.2.6. Segurança no AWS

A segurança na plataforma AWS é definida em várias áreas, incluindo:

- **Responsabilidade Compartilhada do Ambiente:** em geral o cliente do AWS é responsável pelo sistema operacional, incluindo patches e updates de segurança, e pela configuração do grupo de segurança.
- **Responsabilidade e Controle Ambiental:** obediência a uma série de normas, incluindo SAS70 Type II e ISO27001.

- **Utilização de Princípios de Segurança:** o processo de desenvolvimento do AWS inclui utilizar as boas práticas de segurança.
- **Informação e Comunicação:** o AWS utiliza diversos métodos para comunicar a segurança em nível global. Envolve a definição clara das responsabilidades dos funcionários e uma forma de comunicar eventos importantes no tempo adequado.
- **Monitoramento:** os sistemas AWS são monitorados em termos de desempenho e disponibilidade através do uso de ferramentas. Alarmes são configurados com níveis adequados de *thresholds*.
- **Ciclo de Vida do Funcionário:** AWS possui políticas e procedimentos para delimitar o acesso lógico à plataforma e aos servidores de infraestrutura. As políticas definem responsabilidades funcionais pela administração do acesso lógico e da segurança. Envolve política de senha, revisão do acesso e remoção da conta.
- **Segurança Física:** segurança física nos DATACENTERS com segurança de perímetro e políticas e controle de acesso. Envolve detecção de incêndio, energia, clima e temperatura.
- **Backups:** envolve backup dos dados armazenados no S3, SimpleDB ou EBS como parte da operação normal do AWS.
- **Separação Obtida com Regiões e Zonas de Disponibilidade:** dados e instâncias podem ser armazenados em múltiplas regiões geográficas. Cada região é independente. Dentro de cada região, os serviços EC2, EBS e RDS permitem que as instâncias sejam armazenadas em múltiplas zonas.
- **Segurança de Rede:** envolve técnicas para mitigar ataques DDoS, ataques Man in Middle (MITM), *IP spoofing* e varredura de porta.
- **Segurança da Conta de Usuário:** Envolve o AWS Identify and Access Management (AWS IAM) e Key Rotation.
- **Identity and Access Management:** o IAM permite:
 - Criar usuários e grupos para uma conta AWS.
 - Compartilhar recursos entre usuários de uma conta.
 - Definir uma credencial de segurança para cada usuário.
 - Controlar granularmente o acesso de usuários a serviços e recursos.
 - Ter uma única conta para efeito de pagamento para todos os usuários de uma conta AWS.
- **Multi-Factor Authentication (MFA):** envolve uma camada a mais de segurança para o acesso do usuário. Key Rotation trata de recomendações para rotação de pares de chave de segurança e certificados regularmente.

- **Segurança no EC2:** Envolve segurança do SO no servidor, firewall e chamadas via API. O firewall define que os clientes devem explicitar as portas abertas necessárias para o tráfego *inbound*. O tráfego pode ser restrito por protocolo, porta de serviço, endereço IP da fonte ou bloco CIDR. O firewall pode ser configurado para grupos, permitindo que diferentes classes de instâncias tenham diferentes regras.

A **Figura 8-12** considera, por exemplo, a arquitetura em três camadas para uma aplicação instalada no AWS e os aspectos de segurança para cada uma das camadas implementadas. Vale ressaltar que o firewall não é controlado pelo SO e requer um certificado e uma, chave para autorizar a mudança, adicionando um nível extra de segurança. O estado *default* é negar todo o tráfego de entrada.

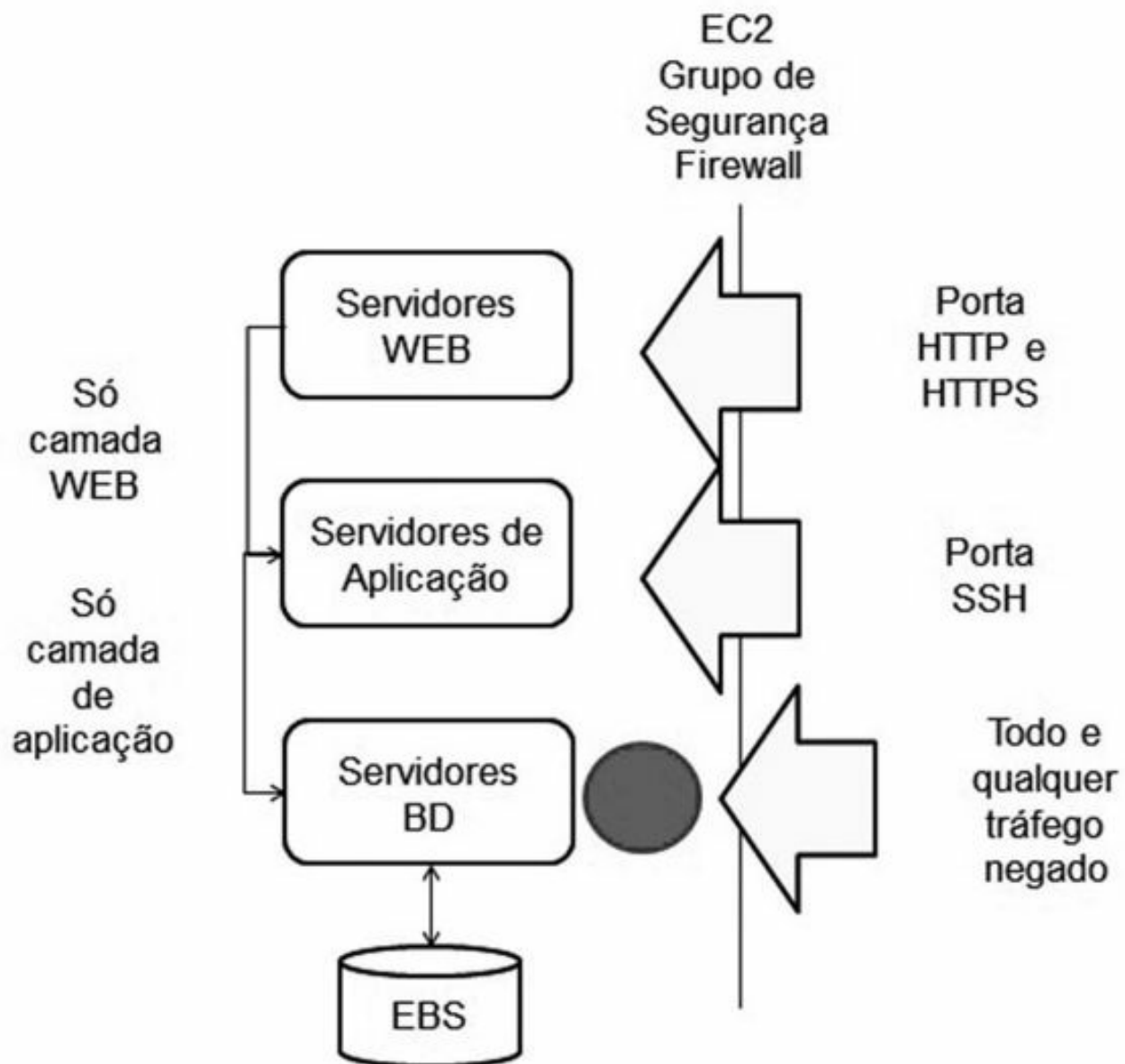


Figura 8-12 – Arquitetura em Camadas do AWS

8.2.7. Capacidade do AWS

O conceito mais importante para entender a capacidade do AWS é a instância. O AWS utiliza a

VIRTUALIZAÇÃO que permite que instâncias de vários usuários rodem no mesmo hardware físico. A VIRTUALIZAÇÃO garante que cada instância lógica receba de forma garantida uma quantidade de memória e tempo de CPU e evita que as instâncias interfiram umas nas outras.

As instâncias podem rodar baseadas em vários sistemas operacionais, incluindo Linux, Windows Server ou OpenSolaris. Quando se inicia uma instância deve-se especificar o tipo. Existem atualmente onze tipos de instâncias, como mostrado na **Tabela 8.3**. A velocidade de cada instância é medida em termos de unidades EC2 Compute, cada unidade sendo equivalente a processadores AMD ou XEON de 1.0 a 1.2 GHz de 2007. A **Tabela 8-3** ilustra os onze tipos de instância do AWS e o preço de referência para instância Linux nos EUA.

Tabela 8-3 – Tipos de Instâncias Amazon

	Nome	Tamanho da CPU (bits)	Cores Virtuais de CPU	Velocidade de Core de CPU (Unidades EC2)	RAM (GB)	Disco Local (GB)	Custo/Hora (\$) EUA/Linux
1	Micro	32 ou 64	1	Até 2	0.613	-	0.02
2	Small	32	1	1	1.7	160	0.085
3	Large	64	2	2	7.5	850	0.34
4	Extra Large	64	4	2	15	1690	0.68
5	High-CPU médio	32	2	2.5	1.7	350	0.17
6	High-CPU extra large	64	8	2.5	7	1690	0.68
7	High-Memory extra large	64	2	3.25	17.1	420	0.50
8	High memory Double Extra Large	64	2	3.25	34.2	850	1.00
9	High Memory Quadruple Extra Large	64	8	3.25	68.4	1690	2.00
10	Cluster Compute	64	8 (2 processadores com 4 cores cada)	33.5 (total)	23	1690	1.60
11	Cluster CPU	64	8(2 processadores com 4 cores cada)+2Nvidia Tesla GPUs	33.5 (total)	22	1690	2.10

8.2.8. Precificação do AWS

O AWS é um serviço pago conforme o uso, existindo um custo separado para cada serviço. A Amazon disponibiliza o *AWS Simple Monthly Calculate* para estimar custos antes de pôr a aplicação em execução.

O AWS cobra pelo uso dos serviços em um nível granular. As dimensões da precificação são:

- Tempo: uma hora de tempo de CPU
- Volume: um gigabyte de dado transferido
- Contagem: número de mensagens enfileiradas
- Tempo e espaço: gigabyte-mês de dado armazenado

Muitos serviços têm mais de uma dimensão de preço. Preços dos serviços tendem a declinar com o tempo devido ao efeito da Lei de Moore e da economia de escala. A precificação também reflete o custo de diferentes regiões.

No portal é possível ter uma posição sobre a conta em termos de consumo de recursos. A Amazon AWS cobra no início de cada mês o consumo de recursos do mês anterior e utiliza o cartão de crédito como opção principal.

A precificação do serviço web EC2 é baseada em várias dimensões e pode variar a qualquer momento. Também varia por região do AWS.

A Amazon AWS também permite aos novos clientes utilizar os serviços EC2 gratuitamente por um ano (*AWS Free Usage Tier*) sob certas condições.

O uso do EC2 é cobrado com base em dimensões que variam com a região e incluem o tipo de instância, dados transferidos, armazenamento de AMI, reservas para endereços IP, armazenamento de dados no EBS e I/O do EBS. Como estes preços mudam com o passar do tempo, melhor consultar o site da Amazon AWS em <http://aws.amazon.com/ec2/pricing/>. Os outros serviços AWS também são precificados e é necessário consultar o site do AWS para saber das últimas atualizações.

8.3. Referências Bibliográficas

Amazon Web Services: Overview of Security Process, Amazon Web Services, May 2011.

Architecting for the Cloud: Best Practices, Amazon Web Services, January 2010, last update, January 2011.

<http://aws.amazon.com>

Barr, Jeff. Host your web site in the clouds: Amazon Web Services made easy. Sitepoint, February 2011.

Building Fault-Tolerant Applications on AWS, Amazon Web Services, May, 2010.

<http://imasters.com.br/artigo/20438/cloud/computacao-em-nuvem-com-o-amazon-web-service> – partes de 01 a 05.

Overview of Amazon Web Services, Amazon Web Services, December 2010.

The Economics of the AWS CLOUD vs. Owned IT Infrastructure, Amazon Web Services. December 2009.

9. Plataforma como Serviço (PaaS)

9.1. Introdução

Os conceitos de SaaS e IaaS já são bem conhecidos e, de alguma forma, já utilizados por diversas organizações. O conceito SaaS trata da troca de um modelo de software baseado em venda de licenças (Capex) por um modelo baseado no uso do software como serviço (Opex). O conceito IaaS trata da utilização da infraestrutura de terceiros como serviço. Os requisitos de banda, latência, poder de processamento são substituídos por garantias do nível de serviço traduzidos por disponibilidade e desempenho adequados.

O conceito de PaaS, menos conhecido, está vinculado ao uso de ferramentas de desenvolvimento de software oferecidas por provedores de serviços, onde os desenvolvedores criam as aplicações e as desenvolvem utilizando a Internet como meio de acesso. Neste caso, o provedor oferta a plataforma de desenvolvimento.

PaaS tem a ver com utilizar uma plataforma de desenvolvimento de terceiros. Na plataforma ofertada rodam os aplicativos e se armazenam os dados. A grande diferença em relação a um modelo convencional de terceirização é que a plataforma roda em DATACENTERS de provedores externos como a Microsoft com seu Windows Azure e é acessada via Internet. Os desenvolvedores estão do outro lado da rede.

Plataforma baseada em nuvem, na visão de David Chappel em seu artigo *The Benefits and Risks of CLOUD Platforms: A Guide for Business Leaders*, é qualquer coisa que pode rodar as aplicações e armazenar os dados, mas que é fornecida por provedores externos e acessadas via Internet. A **Figura 9-1** ilustra a ideia de plataforma.

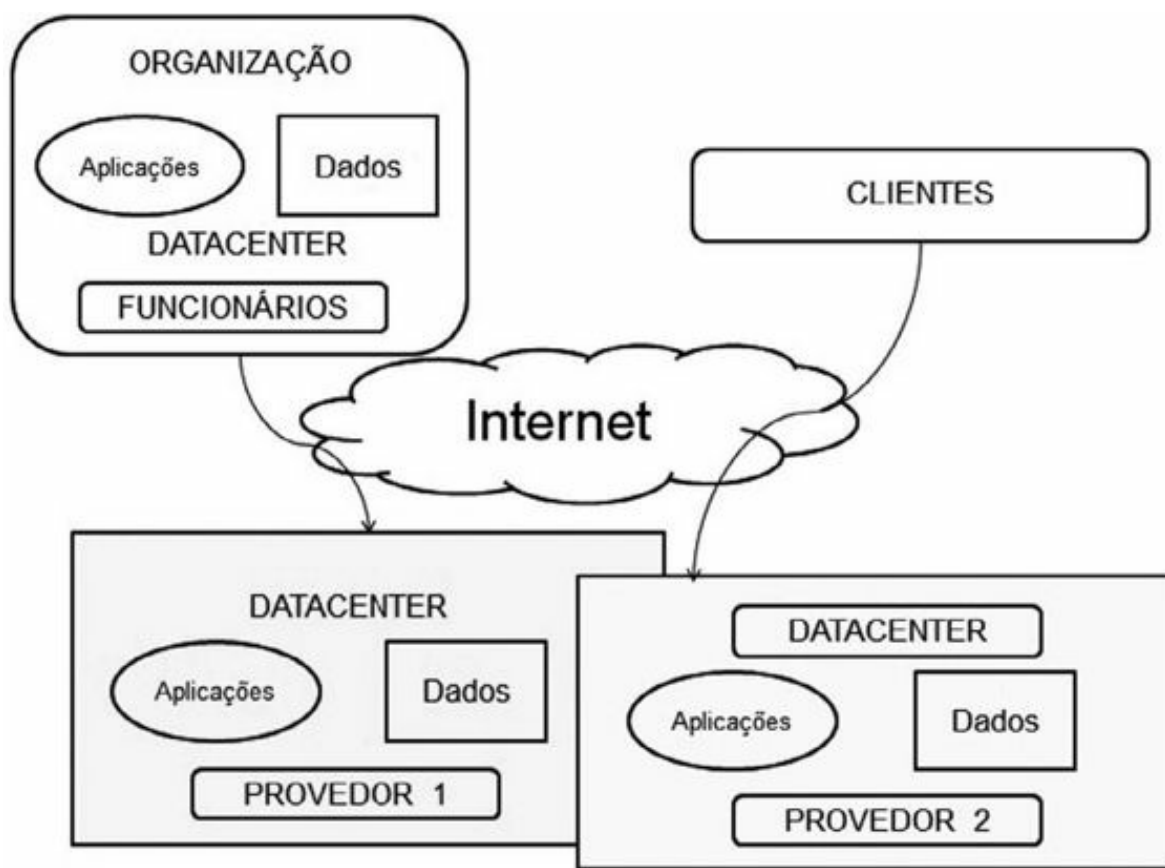


Figura 9-1 – Plataforma de nuvem

Servidores físicos apresentam altos custos de operação devido ao mau aproveitamento dos recursos e gerenciamento complexo. Máquinas ou servidores virtuais surgiram como solução. A possibilidade de fazer várias máquinas virtuais rodarem em um mesmo servidor físico otimiza o uso de recursos, mas também pode causar problemas se as instâncias virtuais são criadas de forma desordenada. Os custos de operação e gerenciamento nesta situação também são altos. A abstração no nível da nuvem permite melhorar o gerenciamento dos ambientes virtualizados. IaaS permite que os usuários aloquem máquinas virtuais através de um portal web. No nível IaaS o que se faz é estruturar e configurar máquinas virtuais, armazenamento e redes. Com PaaS isto já está resolvido.

A Microsoft aposta no modelo PaaS, que, segundo ela:

- Permite manter foco nas aplicações.
- Fornece um ambiente padrão independente das tecnologias utilizadas.
- Mantém o ambiente de software e sistema operacional.
- Propicia serviços prontos para uso que suportam as aplicações.
- Propicia um ambiente *on-demand*.
- Pode ser construído para resistir a falhas.

O modelo PaaS foi inicialmente fornecido por nuvem pública, mas agora também pode ser utilizado em nuvem privada. No caso da Microsoft, o Windows Azure pode ser utilizado tanto por provedores regionais como dentro da organização. Para expandir a plataforma PaaS a Microsoft

fornece o Windows Azure Platform Appliance (ITPAC) também para nuvens privadas. Este *appliance* pré-configurado permite que a plataforma Windows Azure possa ser introduzida em qualquer DATACENTER de uma nuvem, quer seja pública ou privada.

O Windows Azure da Microsoft foi lançado em 2008 e, mesmo sendo classificado como um modelo PaaS, mantém muitas características de um modelo IaaS. Na verdade, o Windows Azure é uma plataforma que combina IaaS com PaaS e possui também grande integração com as tecnologias *on-premises* da Microsoft.

9.2. Conceito na Prática: Windows Azure

9.2.1. Introdução

A plataforma Windows Azure roda atualmente em seis DATACENTERS (dois nos EUA, dois na Europa e dois na Ásia). O desenvolvedor escolhe onde deve rodar a aplicação. Se a mesma aplicação roda em mais de um DATACENTER, o *Windows Azure Traffic Manager* pode distribuir a aplicação com base em melhor desempenho, alta disponibilidade ou balanceamento de carga. A **Figura 9-2** ilustra o uso do *Traffic Manager*.

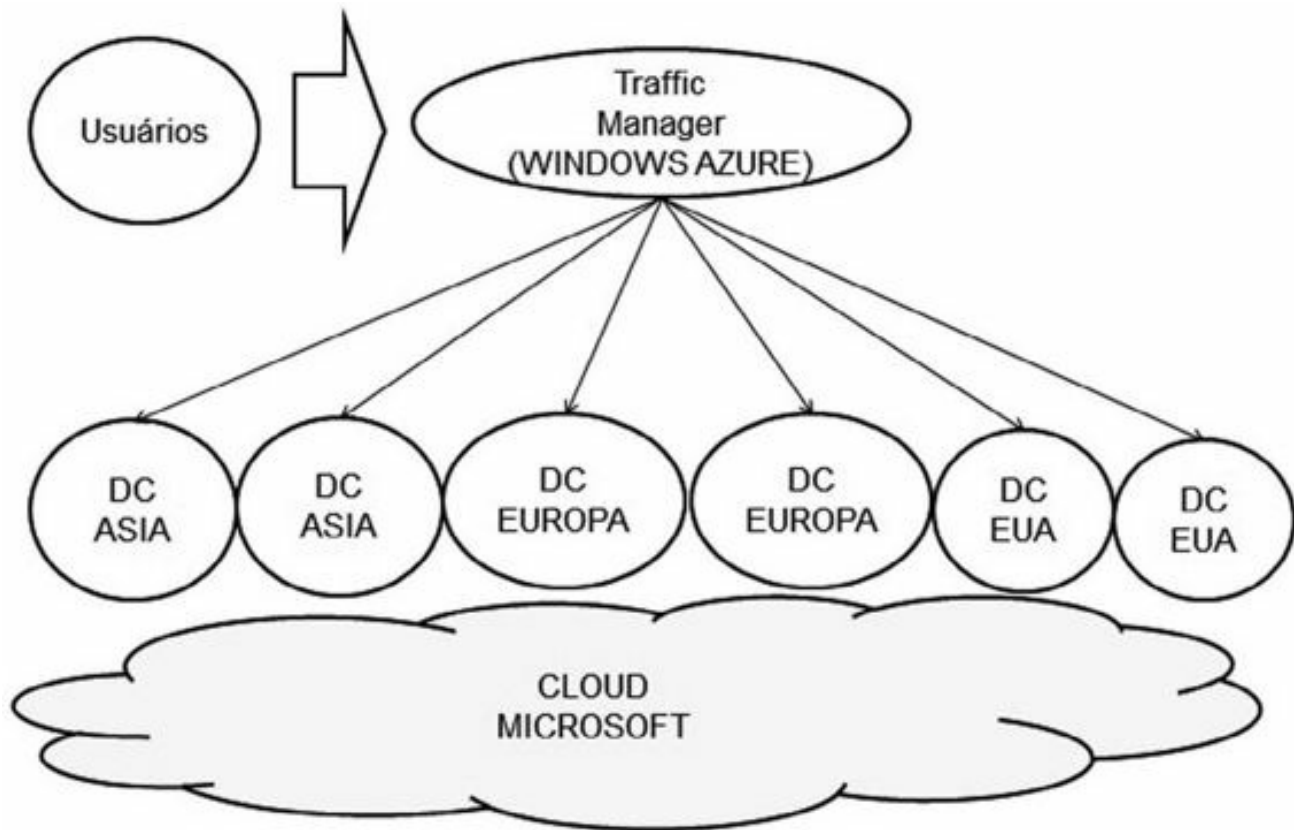


Figura 9-2 – Traffic Manager

Aplicações locais sempre são construídas em algum tipo de plataforma que inclui um sistema operacional, uma forma de armazenamento de dados e outros componentes. As aplicações executadas na nuvem precisam de uma plataforma similar.

O Windows Azure entrega uma plataforma de serviços para desenvolvedores de aplicação. A plataforma Windows Azure, ilustrada na [Figura 9-3](#), contém quatro partes:

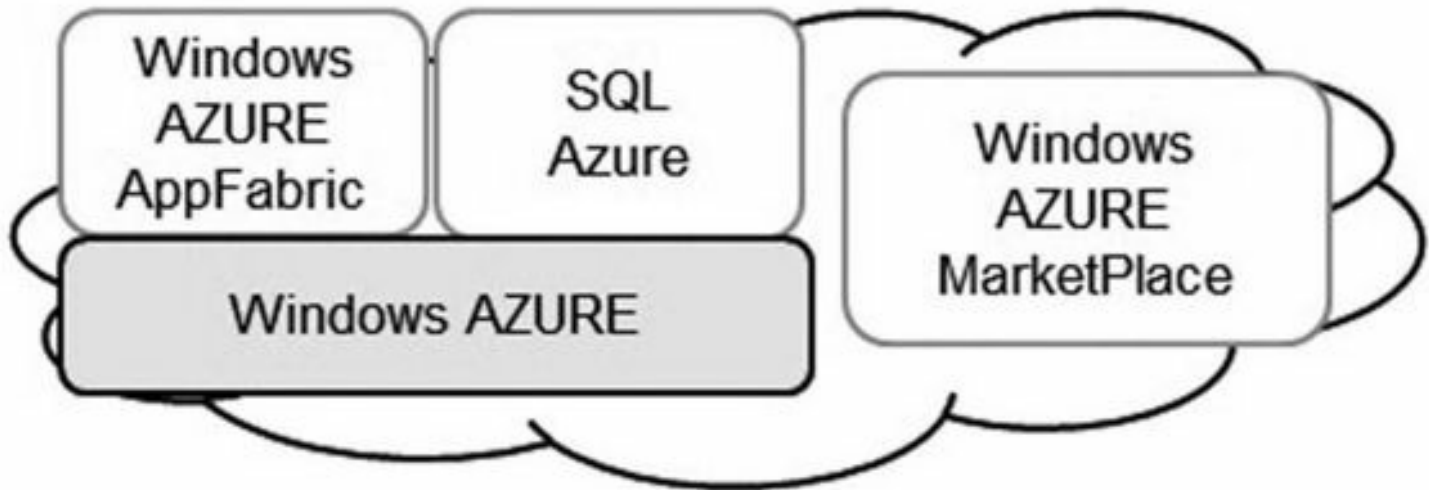


Figura 9-3 – Plataforma Windows Azure

- **Windows Azure:** ambiente Windows para rodar aplicações e armazenar dados em servidores de DATACENTERS Microsoft.
- **SQL Azure:** serviços de dados relacionais na nuvem baseada em SQL Server.
- **Windows Azure AppFabric:** serviços de infraestrutura para aplicações rodando na nuvem ou *on-premises*. Agrupa recursos para hospedagem e execução de serviços na nuvem, com controle de acesso e segurança. Entre os principais grupos de capacidades do Windows Azure AppFabric tem-se: barramento de serviço, controle de acesso e caching.
- **Windows Azure Marketplace:** Serviços online para comprar dados e aplicações baseados na nuvem. O marketplace fornece um único mercado para que vários provedores de conteúdo disponibilizem seus dados para venda por meio de algumas ofertas diferentes (cada oferta pode disponibilizar um subconjunto ou uma exibição diferente dos dados, ou pode disponibilizá-los com diferentes termos de uso). Os provedores de conteúdo especificam os detalhes de suas ofertas, inclusive os termos de uso que regem uma compra e o modelo de precificação.

9.2.2. Windows Azure

O objetivo do Microsoft Windows Azure é fornecer uma plataforma que pode ser usada para a execução de aplicações Windows e o armazenamento de dados na nuvem. Desktops de empresas e de consumidores se relacionam com aplicações e dados armazenados em servidores colocados na nuvem. O relacionamento agora passa a ser ditado por aquisição de serviços e não por aquisição de licença de software.

O Windows Azure é executado em máquinas presentes nos DATACENTERS da Microsoft. Em

vez de fornecer o software para que os clientes instalem e executem as aplicações em seus próprios computadores, o Windows Azure é um serviço: os clientes o utilizam para executar aplicações e armazenar dados em máquinas que pertencem à Microsoft. Essas aplicações podem fornecer serviços para empresas, consumidores ou ambos. Alguns exemplos de aplicações que podem ser construídas no Windows Azure são:

- **Aplicação SaaS voltada para usuários corporativos utilizando SaaS** (*Software as a Service* – Software como Serviço). Os Internet Software Vendors (ISVs) podem usar o Windows Azure como base para várias aplicações SaaS orientadas aos negócios.
- **Aplicação SaaS voltada para os consumidores.** O Windows Azure foi projetado para dar suporte a softwares escalonáveis, portanto, uma empresa que pretenda atingir um grande mercado de consumidores pode escolhê-lo como base para uma nova aplicação.

As empresas podem usar o Windows Azure para construir e executar aplicações que serão usadas pelos próprios funcionários. Embora essa situação não exija a escalabilidade de uma aplicação voltada para os consumidores, a confiabilidade e o gerenciamento oferecidos pelo Windows Azure podem fazer dela uma opção bastante interessante.

Seja qual for a finalidade da aplicação construída no Windows Azure, o sistema operacional oferece os mesmos componentes básicos, como mostra a **Figura 9-4**.

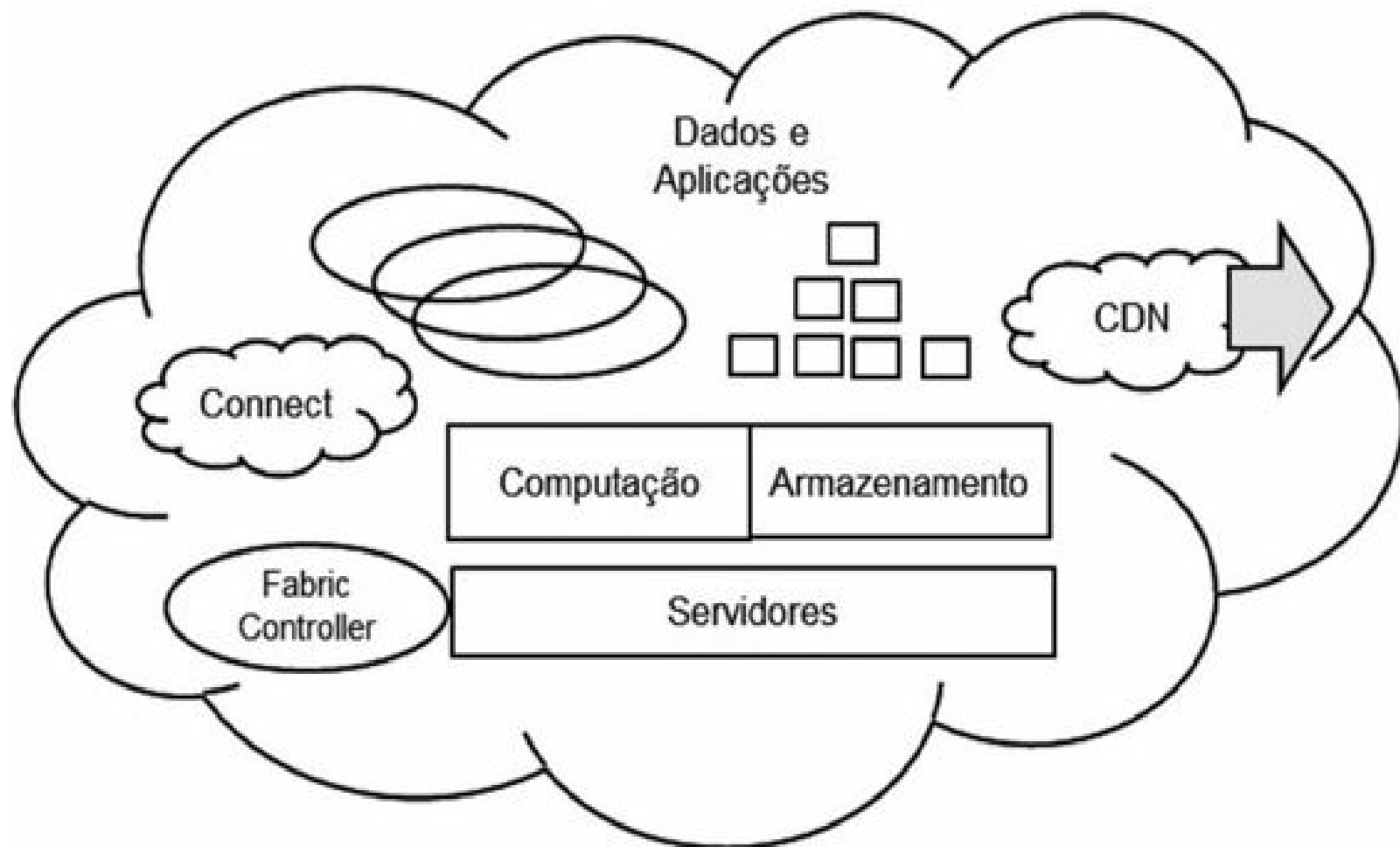


Figura 9-4 – Componentes do Windows Azure

Como os próprios nomes sugerem, o serviço de computação executa aplicações e o serviço de armazenamento armazena os dados. O terceiro componente, o *Fabric Controller* do Windows Azure, oferece uma maneira de gerenciar e monitorar as aplicações que usam essa plataforma na nuvem.

A [Tabela 9-1](#) ilustra as diferenças entre a plataforma Windows e a Plataforma Azure.

Tabela 9-1 – Plataformas Windows Server e o Windows Azure

	Plataforma Windows Server	Plataforma Windows Azure
Sistema Operacional	Windows Server	Windows Azure
DBMS Relacional	SQL Server	SQL Server
Configuração de Hardware	Especificação do cliente	Especificação da Microsoft
Modelo de Update	Cliente aplica os updates	Microsoft aplica os updates
Modelo de Negócio	Compra de Servidores e Licenças	Subscrição de um Serviço
Infraestrutura de Aplicação	.NET, AppFabric, PHP, Java, outros	.NET, AppFabric, PHP, Java, outros
Ferramenta de Desenvolvimento	Visual Studio	Visual Studio
Gerenciamento	System Center	System Center
Identidade	AD	AD
Hypervisor	Hyper-V	Hyper-V

Os clientes podem escolher entre as duas plataformas da Microsoft, ou mesmo utilizar ambas, movendo-se entre as opções *on-premise* e *off-premise*. A [Tabela 9-2](#) ilustra as opções das duas plataformas.

Tabela 9-2 – Opções para clientes Microsoft

	Plataforma Windows Server	Plataforma Windows Azure
DATACENTER Tradicional (TDC) Empresarial ou Provedor Regional	Computação Tradicional, Hardware Tradicional	
Cloud Privada Empresarial ou Provedor Regional	IaaS, Hardware Tradicional	PaaS, Windows Azure Platform Appliance
Cloud Pública Provedor Regional	IaaS, Hardware Tradicional	PaaS, Windows Azure Platform Appliance
Cloud Pública Microsoft		PaaS, GFS

Os servidores são componentes críticos da infraestrutura virtual. Os servidores host, executando o Windows Server 2008 R2 com Hyper-V, oferecem a fundação para a execução de máquinas virtuais convidadas e também oferecem a interface de gerenciamento entre os convidados e o Microsoft System Center Virtual Machine Manager.

Ao consolidar várias funções de servidor em ambientes virtualizados executados em uma única máquina física, a utilização do hardware passa a ser mais efetiva, como também revela possíveis benefícios da Infraestrutura como um Serviço, permitindo dimensionar a infraestrutura com rapidez, adicionando recursos virtuais para receber novas cargas de trabalho ou atender uma demanda maior sempre que necessário.

A principal ferramenta para gerenciar a infraestrutura de nuvem privada Microsoft é o System Center Virtual Machine Manager (VMM). O System Center Virtual Machine Manager pode ser dimensionado em uma ampla variedade de ambientes virtuais, desde um único servidor para ambientes menores até um ambiente empresarial totalmente distribuído que gerencie centenas de hosts executando milhares de máquinas virtuais.

Existem alguns benefícios em fazer o gerenciamento da infraestrutura virtualizada com o VMM, incluindo:

- Suporte à VIRTUALIZAÇÃO para máquinas virtuais executadas no Hyper-V do Windows Server 2008, Microsoft Virtual Server e VMware ESX.
- Suporte ponta a ponta para consolidação de servidores físicos em uma infraestrutura virtual.
- O Performance and Resource Optimization (PRO) para gerenciamento dinâmico e responsivo de infraestrutura virtual (exige o System Center Operations Manager).
- O Intelligent Placement de cargas de trabalho virtuais nos servidores host físicos mais adequados.
- Uma biblioteca completa para gerenciar de maneira centralizada os blocos de construção do DATACENTER virtual.

O System Center Virtual Machine Manager oferece uma camada crítica de gerenciamento e controle que é fundamental para a eficiência de um modelo de nuvem privada Microsoft. O VMM não só oferece uma exibição unificada de toda a infraestrutura virtualizada entre várias plataformas de hosts e uma miríade de sistemas operacionais convidados, como também oferece um conjunto de ferramentas para facilitar a entrada de novas cargas de trabalho com rapidez e facilidade. Por exemplo, o assistente de conversão P2V (*Physical-to-Virtual*, física para virtual) incluído no VMM simplifica o processo de conversão de cargas de trabalho físicas existentes para máquinas virtualizadas. E, em conjunto com o System Center Operations Manager, o recurso *Performance and Resource Optimization* oferece realocação dinâmica de cargas de trabalho virtualizadas para garantir o máximo uso dos recursos de hardware físicos.

A funcionalidade de autoatendimento é um componente fundamental para a entrega de TI como um serviço. Usando o Portal de Autoatendimento do Microsoft System Center Virtual Machine Manager, os DATACENTERS empresariais podem oferecer Infraestrutura como um Serviço para unidades de negócios em sua própria organização. O portal de autoatendimento oferece uma forma para grupos de uma empresa gerenciarem suas próprias necessidades de TI, enquanto a organização de infraestrutura centralizada gerencia um pool de recursos físicos (servidores, redes e hardware relacionado).

As unidades de negócios que se inscreverem no sistema do portal de autoatendimento poderão usar o portal para tratar de algumas funções fundamentais. Por exemplo, por meio de formulários padronizados, as unidades de negócios podem solicitar novas infraestruturas ou alterações em componentes de infraestrutura existentes. Cada unidade de negócios pode enviar solicitações ao administrador de infraestrutura. Os formulários padronizados garantem que o administrador de infraestrutura tenha todas as informações necessárias para atender às solicitações sem precisar entrar em contato repetidamente com a unidade de negócios para obter detalhes.

Quando os componentes da infraestrutura de nuvem privada estiverem prontos, pode-se começar a avaliar cargas de trabalho existentes e migrá-las para o novo ambiente virtualizado. Ao identificar os melhores candidatos à conversão P2V, considere a conversão destes tipos de computadores, em ordem de preferência:

- Servidores subutilizados e que não sejam críticos para a empresa. Ao iniciar pelos computadores menos utilizados e que não sejam críticos para a empresa, você poderá aprender o processo P2V correndo um risco relativamente baixo. Servidores web podem ser bons candidatos.
- Servidores com hardware ultrapassado ou sem suporte que necessitem ser substituídos.
- Servidores com baixa utilização e que estejam hospedando aplicativos internos menos críticos.
- Servidores com utilização mais alta e que estejam hospedando aplicativos menos críticos.
- O restante dos servidores subutilizados.

Em geral, aplicativos críticos para a empresa, como servidores de e-mail e bancos de dados, só deverão ser virtualizados para a plataforma com o Hyper-V no sistema operacional Windows Server 2008 (64 bits).

É possível que, mesmo depois de criar o ambiente de nuvem privada, os requisitos da empresa possam ditar a necessidade de uma capacidade computacional maior do que se gostaria de hospedar ou gerenciar sozinho. O Windows Azure pode oferecer recursos computacionais sob demanda hospedados que podem se integrar diretamente à infraestrutura de nuvem privada existente. Os aplicativos comerciais podem ser implantados na Plataforma Windows Azure, adicionando recursos de computação e armazenamento dimensionados para atender à demanda. Usando o perfil VM do Windows Azure, as máquinas virtuais personalizadas – assim como as que foram criadas na nuvem privada – podem até ser hospedadas em recursos do Windows Azure, oferecendo escalabilidade e capacidade adicionais sempre que for necessário.

A proposta do novo System Center da Microsoft é ser o gerenciador de todo o ambiente de nuvem, quer seja baseado em nuvem pública ou nuvem privada. A [Figura 9-5](#) ilustra a proposta.

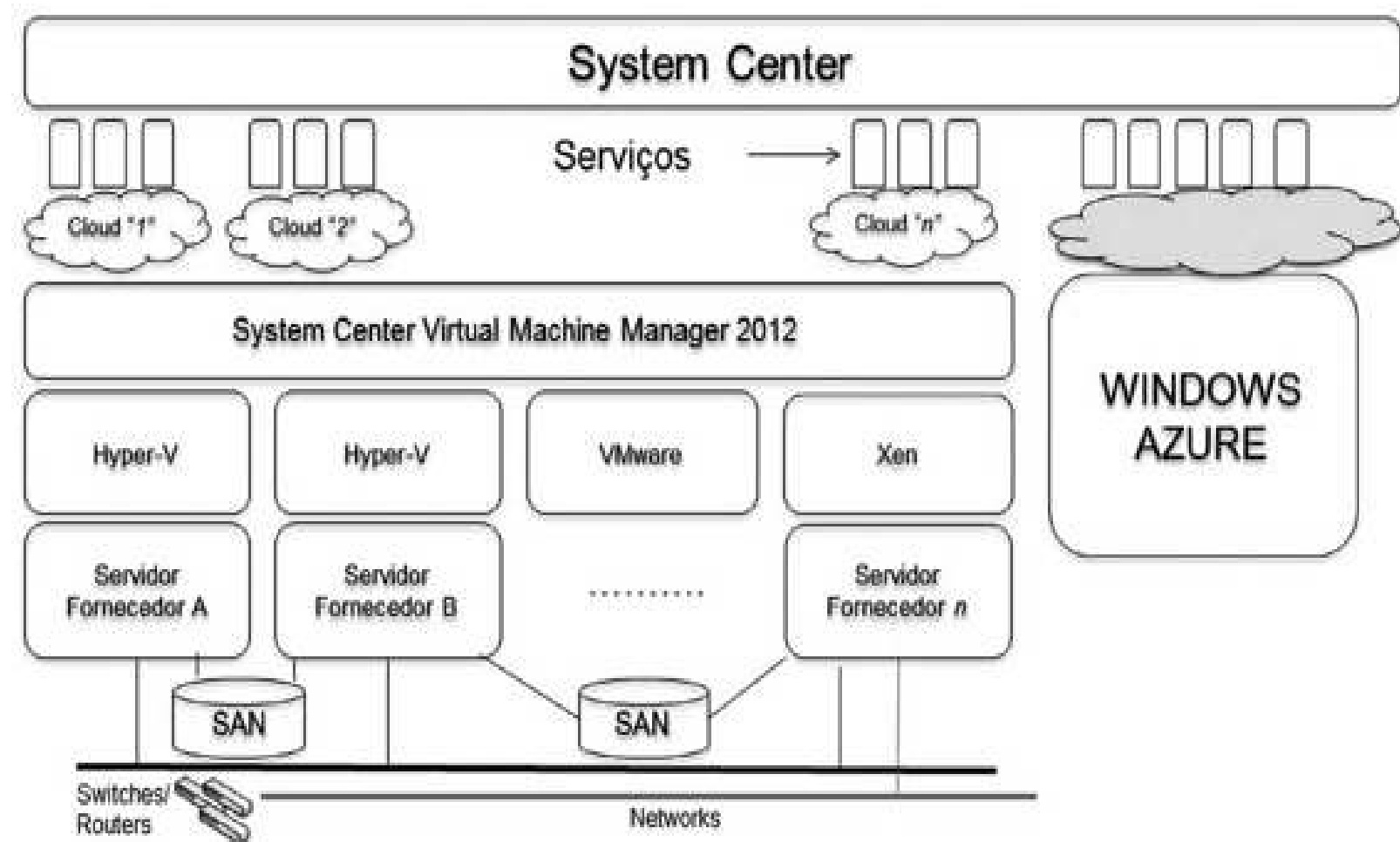


Figura 9-5 – MS System Center (Fonte: Microsoft)

Serviço de Computação

O serviço de computação do Windows Azure pode executar vários tipos de aplicações. O principal objetivo da plataforma, no entanto, é dar suporte a aplicações que têm um grande número de clientes simultâneos. Alcançar esse objetivo escalando verticalmente, ou seja, executando as mesmas aplicações em máquinas cada vez maiores, não é possível. Em vez disso, o Windows Azure foi projetado para dar suporte a aplicações que escalam horizontalmente (*scale-out*), executando múltiplas cópias do mesmo código em vários servidores padrão da indústria.

Para isso, a aplicação Windows Azure pode ter múltiplas instâncias, cada uma executada em sua própria VM (*Virtual Machine* – máquina virtual). Essas VMs são executadas no Windows Server 2008 de 64 bits e fornecidas por um HYPERVISOR (baseado no Hyper-V da Microsoft) que foi modificado para ser usado na nuvem da Microsoft. Para executar uma aplicação, o desenvolvedor acessa o portal do Windows Azure em seu navegador da web utilizando como identificador o Windows Live ID. Depois escolhe se quer criar uma conta de hospedagem para executar as aplicações, uma conta para armazenar dados, ou ambas. Se escolher a conta de hospedagem, poderá então fazer um upload de sua aplicação, especificando o número de instâncias exigido por ela. O Windows Azure cria então as VMs necessárias e executa a aplicação.

É importante notar que o desenvolvedor não pode fornecer sua própria imagem de VM para ser executada pelo Windows Azure. A plataforma fornece e mantém sua própria cópia do Windows. Os desenvolvedores se concentram apenas na criação das aplicações que serão executadas no Windows Azure. As aplicações enxergam o ambiente Windows Server *on-premises* com algumas modificações.

Aplicações construídas no serviço de computação do Windows Azure são estruturadas em um ou mais perfis. As aplicações rodam duas ou mais instâncias de cada perfil, com cada instância rodando na sua própria VM. Desenvolvedores podem escolher três tipos de perfis:

- **Web:** sua intenção é facilitar a criação de aplicações web. Cada instância Web tem IIS 7 pré-configurado e assim permite configurar aplicações [ASP.NET](#), WCF ou utilizando outras tecnologias.
- **Worker:** projetado para rodar uma variedade de código *Windows-based*. Neste caso, o IIS não é pré-configurado. O desenvolvedor pode utilizar tecnologia .NET ou outro software que rode no Windows, incluindo tecnologias não Microsoft.
- **VM:** permite rodar uma imagem Windows Server 2008 R2 provida pelo usuário. Útil para migrar aplicações *on-premises* Windows Server para Windows Azure. VM ainda não salva qualquer mudança feita pelo SO enquanto está rodando.

O desenvolvedor pode submeter a aplicação utilizando o portal Windows Azure e pode usar qualquer combinação de instâncias para os três perfis para criar uma aplicação Windows Azure. Através do portal pode requisitar mais ou menos instâncias para qualquer um dos perfis.

O desenvolvedor pode utilizar qualquer aplicação para Windows como fazia anteriormente, incluindo Java.

A **Figura 9-6** ilustra os perfis para aplicações Windows Azure.

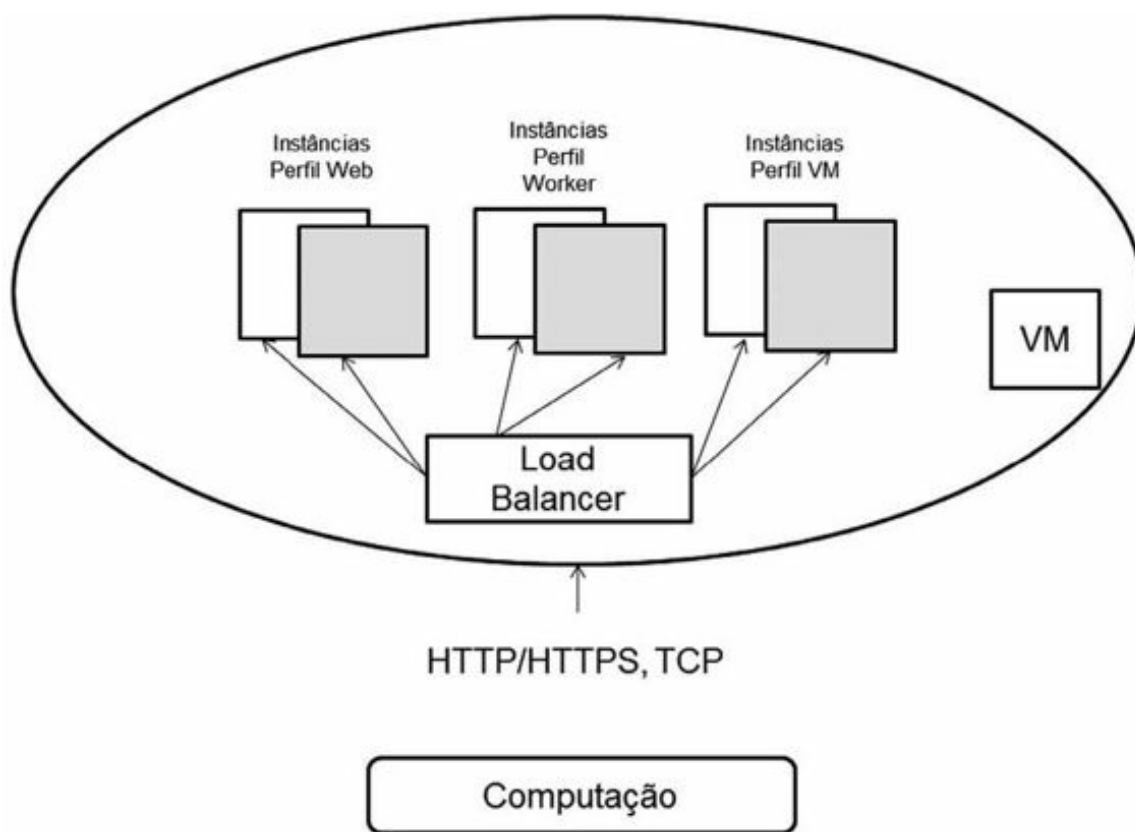


Figura 9-6 – Aplicação Windows Azure

Serviço de Armazenamento

As aplicações trabalham com dados de várias formas diferentes. Por isso, o serviço de armazenamento do Windows Azure oferece três opções: *blobs*, tabelas e filas.

O modo mais simples de armazenar dados no Windows Azure é através dos *blobs*. Um *blob* contém dados binários e uma hierarquia simples: a conta de armazenamento pode ter um ou mais CONTAINERS, e cada um deles possui um ou mais *blobs*. Os *blobs* podem ser grandes – até 50 GB cada – e também podem ter metadados associados.

A [Figura 9-7](#) ilustra os perfis para serviços de armazenamento.

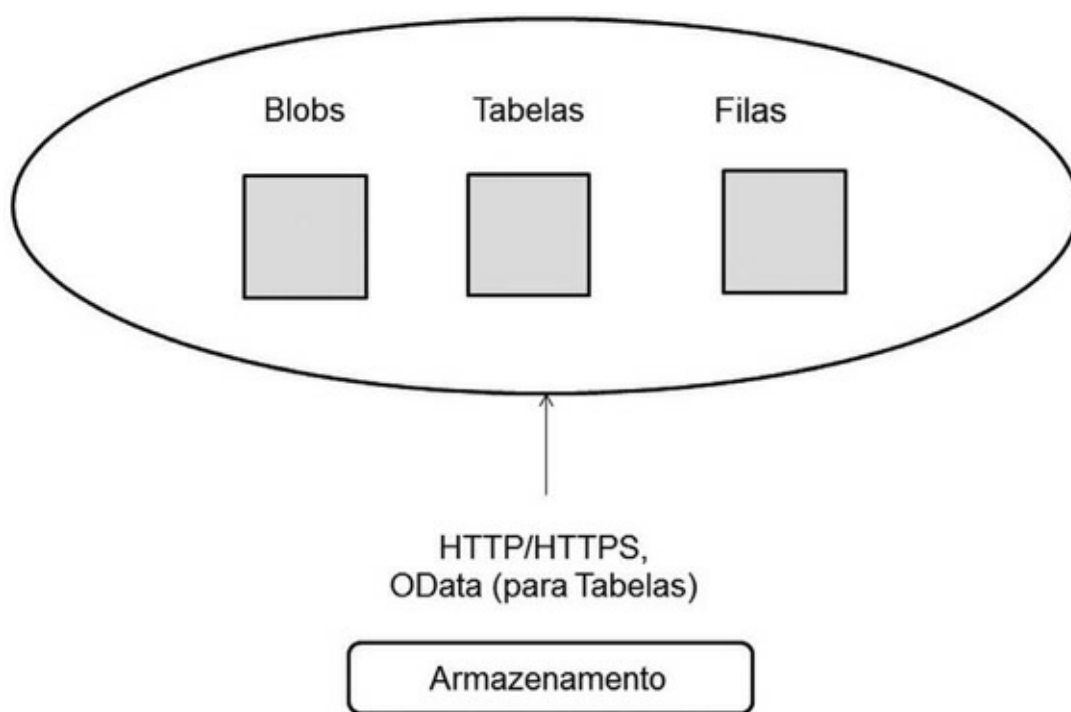


Figura 9-7 – Serviços de Armazenamento no Windows Azure

Os *blobs* são apropriados para algumas situações, mas para outras são muito desestruturados. Para que as aplicações trabalhem com os dados de maneira mais refinada, o armazenamento do Windows Azure oferece tabelas. Essas tabelas não são relacionais. Na verdade, embora sejam chamadas de “tabelas”, seus dados são armazenados em uma hierarquia simples de entidades que contêm propriedades. E em vez de usar o SQL, a aplicação acessa os dados da tabela usando convenções definidas. A razão para o uso desse método é que ele possibilita um armazenamento com escalabilidade horizontal, ou seja, distribui os dados entre várias máquinas – com muito mais eficiência que um banco de dados relacional padrão. Na verdade, uma única tabela do Windows Azure pode conter bilhões de entidades com terabytes de dados.

A principal função dos *blobs* e das tabelas é armazenar e acessar dados. A terceira opção no armazenamento do Windows Azure, as filas, tem um propósito bastante diferente. Por meio das filas, as instâncias da função Web podem se comunicar com as instâncias da função “Trabalhador”, função definida no próprio Azure. Por exemplo, o usuário pode enviar uma solicitação para realizar uma tarefa com uso intensivo de computação através de uma página web implementada por uma função web do Windows Azure. A instância da função web que receber a solicitação poderá gravar uma mensagem em uma fila, descrevendo o trabalho a ser feito. A instância da função “Trabalhador” que estiver servindo essa fila poderá então ler a mensagem e desempenhar a tarefa especificada. Os resultados podem ser retornados através de outra fila ou manipulados de outra maneira.

Blobs, tabelas e filas podem ser úteis. Aplicações que utilizam dados relacionais podem utilizar o SQL Azure, outro componente da plataforma Windows Azure, e podem utilizar acesso SQL padrão.

Fabric Controller

Todas as aplicações Windows Azure e todos os dados do armazenamento do Windows Azure residem em DATACENTERS da Microsoft. Dentro destes DATACENTERS o conjunto de máquinas

dedicadas ao Windows Azure é organizado em um *fabric controller* ou controlador de malha.

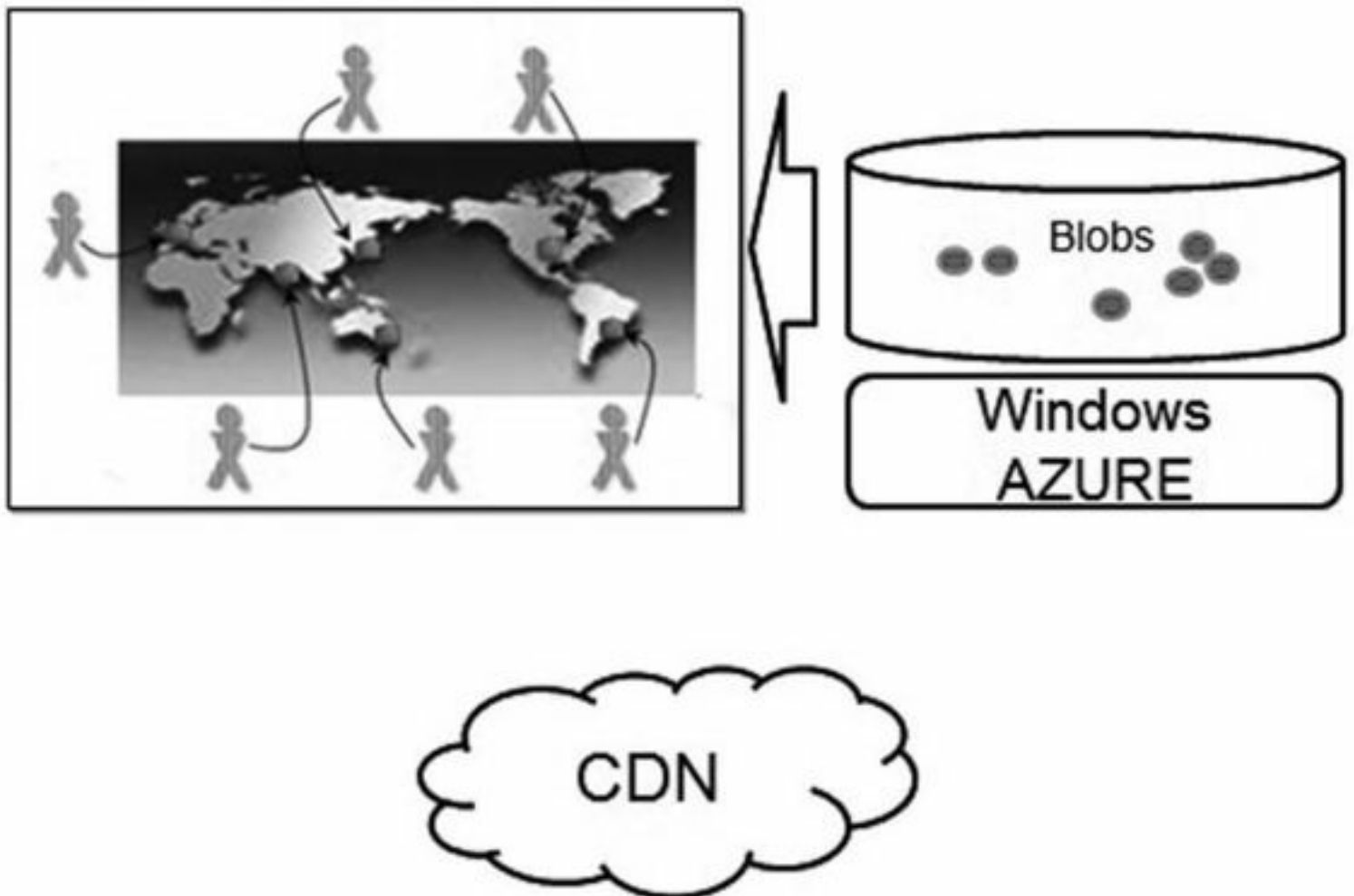
A malha do Windows Azure consiste em um grupo (grande) de máquinas gerenciadas pelo controlador de malha. O controlador de malha é replicado em um grupo de máquinas, e todos os recursos da malha pertencem a ele: computadores, switches, balanceadores de carga, entre outros. Ele pode se comunicar com um agente de malha em cada computador, portanto, ele reconhece todas as aplicações Windows Azure presentes nessa malha (é interessante notar que o controlador de malha vê o armazenamento do Windows Azure apenas como mais uma aplicação, portanto, os detalhes da replicação e do gerenciamento dos dados não ficam visíveis para o controlador).

O controlador de malha monitora todas as aplicações que estão rodando, tendo total controle sobre as VMs e sobre o balanceamento de carga. O controlador de malha, para fazer isto, depende de um arquivo de configuração que é carregado com cada aplicação Windows Azure.

Para alguns perfis de máquina virtual o controlador de malha também gerencia o SO em cada instância. O *fabric* assume que no mínimo duas instâncias de cada perfil estão rodando e assim permite que ele, o controlador, derrube uma instância sem que a aplicação caia inteira.

CDN

Para melhorar a performance o CDN copia os *blobs* em sites mais próximos aos clientes. O CDN funciona como um cache. Na primeira vez que o *blob* é acessado pelo usuário, o CDN armazena uma cópia do *blob* numa localização mais próxima do usuário. No próximo acesso o conteúdo é disponibilizado no CDN. A **Figura 9-8** ilustra a utilização do Windows Azure CDN.



Connect

O Windows Azure Connect é projetado para prover conectividade IP entre aplicações Windows Azure e servidores/desktops rodando fora da *nuvem* Microsoft. Ele requer a instalação de um agente em cada servidor/desktop *on-premises* que se conecta à aplicação Windows Azure. O agente utiliza IPsec para interagir com um perfil particular na aplicação. A [Figura 9-9](#) ilustra o Windows Azure Connect.

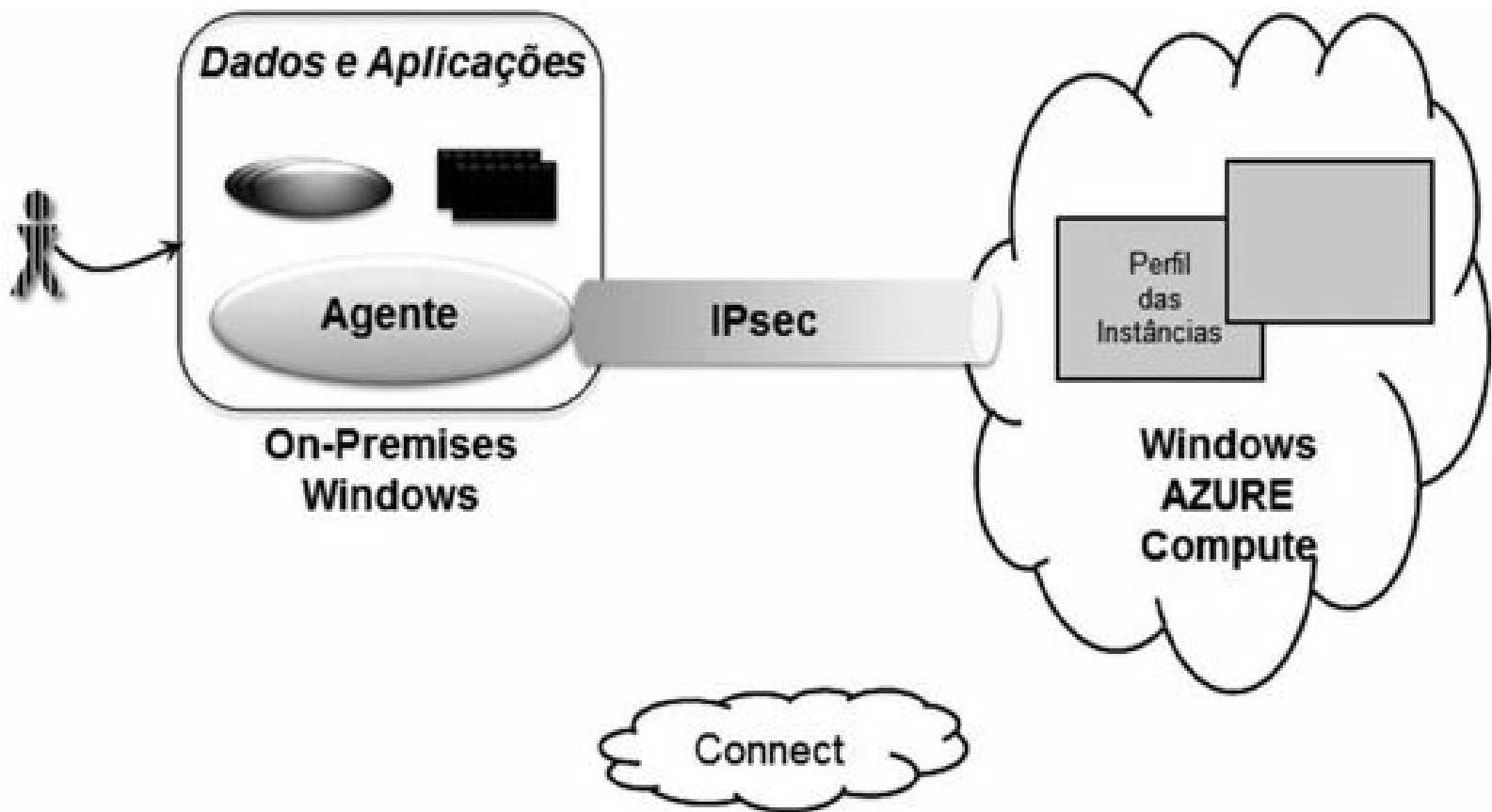


Figura 9-9 – Windows Azure Connect

9.2.3. SQL Azure

O SQL Azure oferece serviços de nuvem para dados relacionais. Possui basicamente três componentes, ilustrados na [Figura 9-10](#).

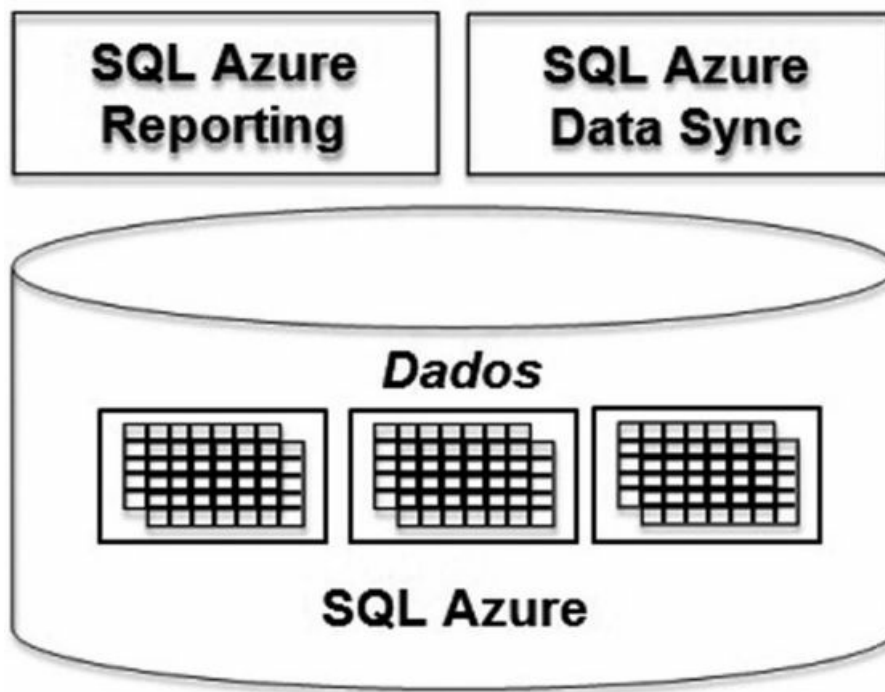


Figura 9-10 – SQL Azure

- **Banco de Dados SQL Azure:** Propicia um sistema de gerenciamento de banco de dados para nuvem. Esta tecnologia permite que aplicações para nuvem e *on-premises* armazenem dados relacionais em DATACENTERS Microsoft. O SQL Azure fornece os serviços de banco de dados com o serviço de nuvem. Uma aplicação rodando o SQL Azure pode rodar sobre o Windows Azure em um DATACENTER corporativo ou até num dispositivo móvel. A aplicação tipicamente acessa os dados via protocolo Tabular Data Stream (TDS). Este é o mesmo protocolo utilizado para acessar um SQL Server local. Assim, um SQL Server Client pode acessar tranquilamente o SQL Azure. Opcionalmente, os dados do SQL Azure podem ser acessados via protocolo ODATA. Cada conta SQL Azure pode ter um ou mais servidores lógicos. Cada servidor pode ter múltiplas base de dados, cada um dos quais com uma base de até 50 GB. Também é possível armazenar um *snapshot* de um SQL Azure em outro servidor, como mecanismo de backup. As funções do SQL Azure são semelhantes aos do SQL Server, mas para alta disponibilidade os dados do SQL Azure são replicados três vezes.
- **SQL Azure Reporting:** é a versão do SQL Server Reporting Services (SSRS) que roda na nuvem. Permite criar e publicar relatórios padrões SSRS de dados na nuvem. O SQL Azure Reporting é projetado para armazenar dados no banco de dados SQL Azure.
- **SQL Azure Data Sync:** Pode ser utilizado para sincronizar dados entre banco de dados SQL Azure, entre banco de dados SQL Azure e um banco de dados SQL Server *on-premise* e até para sincronizar banco de dados SQL Azure em DATACENTERS Microsoft distintos.

9.2.4. Windows Azure AppFabric

O AppFabric propicia serviços básicos de nuvem que podem ser utilizados por aplicações

desenvolvidas para a nuvem ou *on-premises*.

- **Barramento de Serviço:** atua para facilitar a exposição de *end-points* na nuvem que podem ser acessados por outras aplicações *on-premises* ou na nuvem. Cada *end-point* tem uma URL associada onde os clientes podem localizar e acessar o serviço. O *service bus* se encarrega de lidar com o NAT e com os firewalls sem abrir novas portas para as aplicações.
- **Controle de Acesso:** simplifica o acesso a um identificador digital. As opções são muitas e incluem o AD, Windows Live ID, contas do Google e Facebook. O Access Control simplifica esta questão fornecendo suporte para todos estes identificadores. Também fornece um único local para definir perfis, o que cada usuário pode acessar.
- **Caching:** faz a função tradicional de cache para informações frequentemente acessadas, reduzindo a quantidade de vezes que a aplicação faz um *query* no banco de dados.

9.2.5. Windows Azure Marketplace

O Windows Azure Marketplace é uma espécie de mercado para aplicações e dados que podem ser adquiridos na nuvem. O Windows Azure Marketplace possui duas partes:

- **DataMarket:** onde provedores de conteúdo podem fazer datasets disponíveis.
- **AppMarket:** fornece um caminho para criar e expor aplicações de nuvem para potenciais clientes.

9.2.6. Precificação do Windows Azure

Existem diversos componentes na Plataforma Windows Azure. Cada um deles tem seu próprio modelo de cobrança e de contrato de nível de serviço.

Alguns aspectos importantes para entender a precificação:

- **Os custos do Windows Azure Compute são cobrados por hora.** As horas computacionais são calculadas com base no número de horas em que o aplicativo é executado.
- **Existem dois contratos de nível de serviço relacionados à Computação do Windows Azure.** Eles são o SLA de tempo de ativação de Conectividade Mensal e o SLA de tempo de ativação de Instância de Função Mensal. Se o tempo de ativação de Conectividade Mensal ficar abaixo dos 99,95% ou se o tempo de ativação de Instância de Função Mensal ficar abaixo dos 99,9%, um cliente poderia se tornar qualificado ao crédito de serviço somente em relação às cobranças da Computação do Windows Azure.
- **O armazenamento é cobrado em duas taxas.** O armazenamento real é usado a uma taxa por GB. Transações de armazenamento, a cada 10.000 transações. O Armazenamento do Windows Azure é cobrado com base em seu uso médio durante o período de cobrança do armazenamento de *blob*, tabela, fila e unidade. Por exemplo, se você utilizou consistentemente 10 GB de armazenamento na primeira metade do mês e não usou nada

na segunda metade do mês, seria cobrado por seu uso médio de 5 GB de armazenamento.

- **O contrato de nível de serviço de armazenamento do Windows Azure baseia-se em um cálculo simples de “Percentual de Tempo de Atividade Mensal”.** O “Percentual de Tempo de Atividade Mensal” é calculado pela subtração de 100% da Taxa de Erro média para o mês de cobrança para as transações de armazenamento de clientes para uma assinatura individual. A Taxa de Erro é o número total de Transações de Armazenamento com Falha dividido pelo Total de Transações de Armazenamento durante um determinado intervalo (atualmente definido como uma hora). Fatores externos à parte, o contrato de nível de serviço básico para uma transação deve ser processado no período especificado a seguir. A **Tabela 9-3** ilustra os tipos de solicitação no Windows Azure e períodos especificados.

Tabela 9-3 – Solicitações no Windows Azure

Tipo de Solicitação	Período Especificado
PutBlob e GetBlob (inclui blocos e páginas) Obter intervalos de <i>Blob</i> de página válidos	Deve ser concluído no produto de 2 segundos multiplicados pelo número de MBs transferidos no processamento da solicitação
Copiar <i>blob</i>	Deve concluir o processamento em 90 segundos
PutBlockList GetBlockList	Deve concluir o processamento em 60 segundos
Consulta de tabela Operações de lista	Deve concluir o processamento ou retornar uma continuação em 10 segundos
Operações de tabela de lotes	Deve concluir o processamento em 30 segundos
Todas as operações de tabela de entidade única Todas as outras operações de Blob e de mensagem	Deve concluir o processamento em 2 segundos

Transações de armazenamento com falha não incluem as transações que falharam devido a questões não relacionadas a um problema no serviço de armazenamento. Um exemplo deste tipo de transações é a criação de um container. Isso se deve a uma falha na lógica de aplicativo Cliente.

- **A CDN (*Content Delivery Network*) faz parte do serviço de Armazenamento do Windows Azure, mas é cobrada de maneira diferente.** Em sua forma mais simples, as taxas de transferência de dados são determinadas pela região na qual sua solução é implantada. As transferências de dados entre os Serviços do Windows Azure localizados na mesma sub-região não estão sujeitas à cobrança. As transferências entre as sub-regiões são cobradas pelas taxas normais em ambos os lados da transferência. Uma sub-região é o local geográfico de nível mais baixo que você pode selecionar para implantar seus aplicativos e dados associados.

- **O responsável pelo monitoramento de conteúdo é o cliente.** O cliente seleciona um conjunto de agentes da lista padrão do sistema de medição que geralmente estão disponíveis e que representam pelo menos cinco locais geograficamente diversos nas principais áreas metropolitanas do mundo inteiro (excluindo a República Popular da China). O CDN do Windows Azure examinará os dados de qualquer sistema de medição independente aprovado pela Microsoft (atualmente isso inclui Gomez e Keynote), para calcular os níveis de serviço com base em um conjunto de teste. Se os resultados dos testes indicarem um nível de serviço abaixo de 99,9% por mês, o cliente terá direito a um crédito de serviço.
- **O SQL Azure possui dois modelos de cobrança com base no tamanho do banco de dados.** Os dois modelos são Web Edition ou Business Edition. Os bancos de dados de 1 GB ou 5 GB recaem no modelo de preço Web Edition. Um banco de dados de 10 GB, 20 GB, 30 GB, 40 GB ou 50 GB de tamanho recai no modelo Business Edition. Os preços são estabelecidos por mês e se baseiam somente no tamanho do banco de dados escolhido por você e não na quantidade de dados do banco de dados. Há o custo por mês para cada tamanho.

O contrato de nível de serviço para um banco de dados do SQL Azure é um pouco mais complexo do que os de outros componentes.

A fórmula básica é:

$$(Número\ Total\ de\ Minutos\ no\ mês \times Número\ Total\ de\ Usuários - Total\ de\ minutos\ de\ Downtime\ no\ mês) / Número\ Total\ de\ Minutos\ no\ mês \times Número\ Total\ de\ Usuários$$

Assim como acontece com outros contratos de nível de serviço, um nível de serviço abaixo de 99,9% por mês poderia qualificar o cliente a um crédito de serviço.

- **O Windows Azure AppFabric é cobrado com base nos dois serviços atuais, Controle de Acesso e Barramento de Serviço.** Os modelos para ambos são muito simples. O Controle de Acesso se baseia em transações e o Barramento de Serviço se baseia em conexões.

Os dois serviços do AppFabric usam a mesma fórmula para calcular a disponibilidade do serviço.

$$(Número\ Total\ de\ Minutos\ no\ mês \times Número\ Total\ de\ Usuários - Total\ de\ minutos\ de\ Downtime\ no\ mês) / Número\ Total\ de\ Minutos\ no\ mês \times Número\ Total\ de\ Usuários$$

O contrato de nível de serviço baseia-se em dois níveis: nível de serviço e processamento de mensagens. Se um desses serviços ficar abaixo de 99,9% por mês, o cliente poderá se qualificar a um crédito de serviço.

9.2.7. Segurança do Windows Azure

O Windows Azure oferece as dimensões clássicas de segurança da informação.

- **Autenticação mútua de SSL para tráfego de controle interno:** todas as comunicações entre componentes internos do Windows Azure são protegidas com SSL.

- **Gerenciamento de certificados e de chaves privadas:** para diminuir o risco de expor certificados e chaves privadas para desenvolvedores e administradores, eles são instalados por meio de um mecanismo separado do código que os utiliza.
- **Software de cliente de privilégio mínimo:** a execução de aplicativos com privilégio mínimo é amplamente considerada uma prática recomendada de segurança das informações. Por estarem de acordo com o princípio de privilégio mínimo, os clientes não obtêm acesso administrativo a suas máquinas virtuais e, por padrão, o software do cliente no Windows Azure é restrito à execução sob uma conta de baixo privilégio (em versões futuras, os clientes terão a opção de selecionar diferentes modelos de privilégios). Isso reduz o potencial impacto e aumenta a sofisticação necessária de qualquer ataque, exigindo a elevação de privilégios, além de outras explorações. Também protege o serviço do cliente contra ataques de seus próprios usuários finais.
- **Controle de acesso no armazenamento do Windows Azure:** o armazenamento do Windows Azure tem um modelo de controle de acesso simples. Cada assinatura do Windows Azure pode criar uma ou mais Contas de Armazenamento. Cada conta de armazenamento tem uma única chave secreta usada para controlar o acesso a todos os dados naquela Conta de Armazenamento. Isso oferece suporte ao cenário típico no qual o armazenamento é associado a aplicativos e esses aplicativos têm controle total sobre os dados associados.
- **Isolamento do HYPERVISOR, sistema operacional raiz e máquinas virtuais convidadas:** um limite crítico é o isolamento da máquina virtual raiz das máquinas virtuais convidadas e das máquinas virtuais convidadas umas das outras, gerenciadas pelo HYPERVISOR e pelo sistema operacional raiz.
- **Isolamento de controladores de malha:** como orquestradores centrais de grande parte do Windows Azure Fabric, controles significativos foram adotados para minimizar as ameaças a controladores de malha.
- **Filtragem de pacotes:** o HYPERVISOR e o sistema operacional raiz fornecem filtros de pacotes de rede que garantem que as VMs não confiáveis não possam gerar tráfego falsificado, receber tráfego não endereçado a elas, direcionar tráfego para pontos de extremidade de infraestrutura protegidos nem enviar ou receber tráfego de difusão não apropriado.
- **Isolamento de VLAN:** as VLANs são usadas para isolar os FCs e outros dispositivos. As VLANs particionam uma rede de forma que nenhuma comunicação seja possível entre VLANs sem a passagem por um roteador, o que impede que um nó comprometido falsifique tráfego de fora de sua VLAN, exceto para outros nós de sua VLAN, além de não poder bisbilhotar o tráfego que não seja de ou para suas VLANs.
- **Isolamento de acesso de cliente:** os sistemas que gerenciam o acesso a ambientes de cliente são isolados em um aplicativo do Windows Azure operado pela Microsoft. Isso separa, logicamente, a infraestrutura de acesso do cliente de aplicativos e armazenamento do cliente.

- **Exclusão de dados:** quando apropriado, a confidencialidade deverá persistir além do ciclo de vida útil dos dados. O subsistema de armazenamento do Windows Azure indisponibiliza os dados do cliente quando operações de exclusão são chamadas. Todas as operações de armazenamento, incluindo a exclusão, são projetadas para serem instantaneamente consistentes. A execução bem-sucedida de uma operação de exclusão remove todas as referências ao item de dados associado e não pode ser acessada por meio das APIs de armazenamento. Todas as cópias do item de dados excluído são, então, coletadas como lixo. Os bits físicos são substituídos quando o bloco de armazenamento associado é reutilizado para armazenamento de outros dados, como acontece nas unidades de disco rígido de computador padrão.

9.2.8. Capacidade do Windows Azure

A plataforma Windows Azure define e impõe diretivas de forma que os aplicativos executados em uma infraestrutura virtualizada funcionem bem uns com os outros. O reconhecimento dessas diretivas de recurso é importante para a avaliação da capacidade para operações bem-sucedidas e também para a previsão das despesas operacionais para fins de planejamento.

- **Largura de banda da rede:** a largura de banda da rede baseia-se no tamanho da instância de computação solicitada. A [Tabela 9-7](#) a seguir oferece os detalhes dos tamanhos da instância e o compartilhamento de largura de banda para cada um.

Tabela 9-7 – Instância e Largura de Banda de Rede

Instância de computação	CPU (GHz)	Memória	Armazenamento	Largura de banda
XSmall	1	768 MB	20 GB	5 Mbps
Pequena	1,6	1,7 GB	225 GB	100 Mbps
Média	2 x 1,6	3,5 GB	490 GB	200 Mbps
Grande	4 x 1,6	7 GB	1000 GB	400 Mbps
XLarge	8 x 1,6	14 Gb	2040 GB	800 Mbps

- **Alocação de CPU:** assim como a largura de banda de rede, as diretivas de recurso de CPU são implementadas de forma implícita com base no tipo de instância. O número específico de núcleos de CPU que vêm com cada função é mostrado na [Tabela 9-8](#).

Tabela 9-8 – Instância e Alocação de CPU

Instância de computação	CPU garantida
XSmall	Núcleo compartilhado
Pequena	1 Básico

Média	2 Núcleos
Grande	4 Núcleos
XLarge	8 Núcleos

- **Alocação de memória:** a cada instância é provisionado um valor de memória pré-configurado. As instâncias de função obtêm suas alocações de memória com base no tipo de função, da memória restante no servidor físico, depois que o sistema operacional raiz obtém sua cota. A [Tabela 9-9](#) ilustra a alocação de memória.

Tabela 9-9 – Instância e Alocação de Memória

Instância de computação	Memória Garantida
XSmall	0,768 GB
Pequena	1,750 GB
Média	3,50 GB
Grande	7,00 GB
XLarge	14,0 GB

- **Armazenamento de instância:** Cada instância do Azure obtém uma alocação de armazenamento de disco volátil. Dependendo de seu aplicativo ter ou não monitoração do estado, você precisa ter duas considerações principais. Se os aplicativos armazenam informações neste repositório, se a configuração cleanOnRoleRecycle não estiver definida como “falso”, o repositório será reciclado durante uma reinicialização da instância. O repositório também não persistirá se a instância tiver de ser realocada para outro servidor físico. Isso pode ocorrer se houver um problema subjacente no servidor físico, uma vez que o Windows Azure manterá automaticamente as instâncias solicitadas. A [Tabela 9-10](#) ilustra o armazenamento em disco.

Tabela 9-10 – Instância e Armazenamento em Disco

Instância de computação	Armazenamento em disco
XSmall	20 GB
Pequena	220 GB
Média	490 GB
Grande	1000 GB
XLarge	2040

- **Armazenamento do Windows Azure:** ao contrário do armazenamento de instância, o

Armazenamento do Windows Azure persiste. O armazenamento do Windows Azure oferece quatro abstrações de objeto para armazenamento de dados.

- **Blobs:** oferecem uma interface simples para o armazenamento de arquivos nomeados junto com os metadados do arquivo.
 - **Tabelas:** oferecem armazenamento estruturado massivamente escalonável. Uma Tabela é um conjunto de entidades que contém um conjunto de propriedades. Um aplicativo pode manipular as entidades e a consulta sobre qualquer uma das propriedades armazenadas em uma Tabela.
 - **Filas:** oferecem armazenamento e entrega confiáveis de mensagens para um aplicativo para criar um fluxo de trabalho escalonável e menos rígido entre as diferentes partes (funções) de seu aplicativo.
 - **Unidades:** oferecem volumes NTFS duráveis para que os aplicativos do Windows Azure os usem. Isso permite que os aplicativos usem APIs do NTFS existentes para acessar uma unidade durável anexada à rede. Cada unidade é um blob de página anexado à rede formatado com um único volume VHD NTFS. Nesta postagem, não nos concentraremos nas unidades, uma vez que a escalabilidade deles é de um único *blob*.
- **Conta de armazenamento:** ao criar uma conta de armazenamento, pode-se especificar um local para hospedar os dados de armazenamento. Os seis locais oferecidos atualmente são:
- Centro-norte dos EUA
 - Centro-sul dos EUA
 - Norte da Europa
 - Europa Ocidental
 - Leste da Ásia
 - Sudeste da Ásia

Como prática recomendada, deve-se escolher o mesmo local para a conta de armazenamento e instâncias.

- **Capacidades de Armazenamento:** uma única conta de armazenamento pode ter um tamanho de até 100 TB.
- Os *blobs* têm duas versões: de bloco e de tabela. Um *blob* de bloco pode ter um tamanho de até 200 GB. Um *blob* de tabela pode ter até 1 TB.
 - Uma tabela é um agrupamento de Entidades (Linhas). Uma Entidade é composta de Propriedades. Uma única entidade pode armazenar até 255 propriedades para as quais o tamanho combinado de todas as propriedades em uma entidade não pode exceder 1 MB.
 - Uma Fila contém Mensagens. Não há limite para o número de Mensagens com suporte de uma Fila além dos limites da conta de Armazenamento. Uma mensagem em

uma Fila não pode ter mais de 8 KB.

- As unidades são manipuladas como *blobs* de página e podem um tamanho de até 1 TB.

9.2.9. Windows Azure versus Amazon AWS

É inevitável fazer uma comparação entre as duas plataformas, mesmo sabendo que os fabricantes as posicionam de forma diferente. Na verdade o Windows Azure possui muitas características de Infraestrutura como Serviço (IaaS), apesar de ser posicionado como Plataforma como Serviço (PaaS).

Por sua vez, os serviços AWS da Amazon estão claramente posicionados como Infraestrutura como Serviço (IaaS). A ideia do AWS é entregar a infraestrutura como serviço para os programadores, o que a equipe da Amazon chama de DATACENTER programável. Os programadores, por sua vez, devem construir as aplicações para utilizar os recursos disponibilizados pela infraestrutura provendo desempenho e disponibilidade para os usuários, tudo baseado em serviços web.

Uma vantagem do AWS é ter uma oferta de virtualização para tecnologias Microsoft e não Microsoft. AWS, por exemplo, oferece, além do MySQL, os gerenciadores de banco de dados Oracle e SQL Server.

O AWS também oferece plataformas Windows e Linux, enquanto o Azure oferece só a família Microsoft baseada no Windows Server e SQL Server. Por outro lado, para instalações que são baseadas em Microsoft, a plataforma Azure é muito bem integrada, incluindo o gerenciamento através da família System Manager. No Windows Azure o desenvolvedor, mesmo utilizando Ruby ou PHP, dependerá do Visual Studio quando for fazer o *deployment da aplicação*.

Tanto o AWS como o Azure oferecem *trails*, mas com condições e contratos diferentes. Sugere-se a consulta aos sites dos fabricantes para verificar as condições atuais.

As instâncias para o AWS acontecem em maior número e as opções para cluster e para grandes memórias (High Memory) e grandes CPUs (High CPU) são o diferencial da oferta quando comparado com o Azure. De qualquer forma, as instâncias para as duas plataformas são segmentadas de forma próxima. Para instâncias de tamanho próximos baseadas em Windows os preços são equivalentes. As instâncias Linux do AWS são normalmente 25% a 30% mais baratas que as instâncias Windows.

A **Tabela 9-4** compara as principais funcionalidades das duas plataformas.

Tabela 9-4 – Funcionalidades Windows Azure versus Amazon AWS

Funcionalidade	Azure	AWS
Computação		
Virtualização	Windows Azure perfis Web, Worker e VM	EC2
Disponibilização de Conteúdo		
Serviço de Disponibilização de Conteúdo	CDN	CloudFront
Identidade		
Gerenciamento de Indentidade	Windows Azure AppFabric Access Control Service	Serviços IAM
Rede		
Gerenciamento do Tráfego	Traffic Manager	Route 53
Redes Privadas	Connect	Direct Connect
Armazenamento		
Armazenamento de Dados Binários	Blob	S3
NoSQL	Serviço de Tabelas	SimpleDB
Banco de Dados Relacional	SQL Azure	RDS
Mensageria		
Queuing	Filas AppFabric	SQS
Notificações	AppFabric Topics	SNS
Caching	AppFabric Caching	ElastiCache

9.3. Referências Bibliográficas

<http://technet.microsoft.com/pt-br/cloud/default>

Introducing the Windows Azure Platform, David Chappel, outubro de 2010.

Introducing Windows Azure, David Chappel, outubro de 2010.

10. Software como Serviço (SaaS)

10.1. Introdução

Software como um serviço (*Software as a Service* – *SaaS*) é uma modalidade de serviços de nuvem onde os aplicativos de interesse para uma grande quantidade de usuários passam a ser hospedados na nuvem como uma alternativa ao processamento e armazenamento local. Os aplicativos são oferecidos como serviços por provedores e acessados pelos clientes por aplicações como o browser. Todo o controle e gerenciamento da rede, sistemas operacionais, servidores e armazenamento é feito pelo provedor de serviço. O Google Apps e o [SalesForce.com](https://www.salesforce.com) são exemplos de SaaS.

O SaaS é uma espécie de evolução do conceito de ASPs (*Application Service Providers*), que forneciam aplicativos “empacotados” aos usuários de negócios pela Internet. Havia, de certa forma, nessas tentativas iniciais de software entregue pela Internet, mais pontos em comum com os aplicativos tradicionais *on-premise* (instalados no local), como licenciamento e arquitetura, do que com a proposta dos novos aplicativos baseado em SaaS. Os aplicativos baseados em ASP foram originalmente construídos para serem aplicativos de um único inquilino; sua capacidade de compartilhar dados e processos com outros aplicativos é limitada e oferece poucos benefícios econômicos em relação aos seus similares instalados no local.

Atualmente, espera-se que os aplicativos SaaS aproveitem os benefícios da centralização através de uma arquitetura de instância única, para vários inquilinos, e para oferecer uma experiência rica em recursos, que compete com os aplicativos *on-premise* de mesmo tipo. Aplicativos SaaS típicos são oferecidos diretamente pelo fornecedor, ou por intermediários, que reúnem ofertas SaaS de vários fornecedores, oferecendo-as como parte de uma plataforma de aplicativos unificada.

Diferentemente do modelo de licenciamento único, normalmente usado para software instalado no local, o acesso ao aplicativo SaaS é quase sempre vendido de acordo com um modelo de assinatura: os clientes pagam uma taxa contínua para uso do aplicativo. Os planos de cobrança variam de acordo com o aplicativo; alguns provedores cobram taxa fixa para acesso ilimitado a alguns recursos do aplicativo, ou para todos; outros cobram taxas variáveis baseadas no uso.

O provedor de SaaS hospeda o aplicativo, os dados e implanta patches e atualizações do aplicativo de modo centralizado e transparente, possibilitando o acesso aos usuários finais pela Internet, via navegador ou aplicativo *smart-client*. Muitos fornecedores oferecem interfaces de programação de aplicativo (APIs) que expõem os dados e a funcionalidade dos aplicativos aos desenvolvedores para uso na criação de aplicativos compostos. Vários mecanismos de segurança podem ser usados para manter a segurança de dados sigilosos, na transmissão e no armazenamento. Os provedores de aplicativos podem fornecer ferramentas que permitem aos clientes modificar o esquema de dados, o fluxo de trabalho e outros aspectos operacionais do aplicativo, de acordo com o

seu uso.

Um tema importante na utilização do modelo SaaS é a migração de aplicações que rodam em DATACENTERS corporativos para nuvens públicas. Existem muitos aspectos a serem considerados para realizar a integração entre nuvens públicas e DATACENTERS corporativos. Segundo a Oracle, existem cinco principais desafios:

- Organizações devem ser capazes de carregar dados rapidamente para permitir que as aplicações de nuvens rodem em um tempo adequado.
- Os dados devem ser mantidos sincronizados seguidamente em tempo real.
- Processos devem ser totalmente integrados, mesmo considerando que podem depender de aplicações que rodam na nuvem pública e no DATACENTER privado.
- Contas de usuário e privilégios de acesso para nuvens públicas devem ser gerenciados.
- Informação sensível deve permanecer segura no novo ambiente híbrido.

10.2. Benefícios do SaaS

O SaaS oferece oportunidades substanciais para que empresas, de todos os portes, deixem de enfrentar os riscos da aquisição de software e transfiram o departamento de TI de um centro de custo para um centro de valor.

10.2.1. Melhor Gerenciamento dos Riscos da Aquisição de Software

Tradicionalmente, tem sido uma grande responsabilidade a implantação de sistemas de software de larga escala, críticos para o negócio, como os pacotes de aplicativos de ERP e CRM. A implantação desses sistemas em uma empresa de grande porte pode custar centenas de milhares de dólares, pagos no primeiro momento, no custo do licenciamento e infraestrutura, e, em geral, exige um exército de profissionais e consultores de TI para a personalização e a integração com os outros sistemas e dados da organização. As exigências de tempo, equipe e orçamento de uma implantação dessa magnitude representam um risco significativo para empresas de qualquer porte e, frequentemente, fazem com que esse software fique fora do alcance de empresas menores, que poderiam, se assim não fosse, obter dele grande proveito em todos os sentidos.

Os aplicativos SaaS não exigem a implantação de grande infraestrutura no local do cliente, o que elimina ou reduz, drasticamente, o compromisso de recursos adiantados. Sem investimento inicial significativo para amortizar, a empresa que implanta um aplicativo SaaS que resulte em produção de resultados desalentadores poderá desistir e caminhar em outra direção, sem ter que abandonar a cara infraestrutura de suas instalações.

Além disso, se a integração personalizada não for necessária, os aplicativos SaaS podem ser planejados e executados com um mínimo de empenho e de atividades de distribuição. Isso também possibilita que vários fornecedores SaaS ofereçam “test drives” com baixo risco de seus produtos, por um período limitado, como 30 dias. Esse procedimento possibilita aos clientes empresariais uma oportunidade de testar o software antes de comprá-lo e ajuda a eliminar a maior parte do risco que

cerca a compra de um novo produto.

10.2.2. Mudança no Foco da TI

Com o SaaS, o trabalho de implantar um aplicativo e mantê-lo em funcionamento, dia após dia – testando e instalando patches, gerenciando atualizações, monitorando o desempenho, assegurando alta disponibilidade, etc. – ficará sob a responsabilidade do provedor. Ao transferir a responsabilidade dessas atividades de “sobrecarga” a terceiros, o departamento de TI poderá se concentrar melhor em atividades de valor mais elevado que se alinham e dão suporte às metas comerciais da empresa. Em lugar de ser principalmente reativo e operacional, o CIO e a equipe de TI poderão trabalhar com maior eficiência como estrategistas de TI para o restante da empresa, trabalhando com unidades do negócio para entender suas necessidades e fazer recomendações sobre como melhor usar a tecnologia para alcançar seus objetivos.

10.3. Diferenças entre Software Convencional e SaaS

Na forma “pura” do SaaS, o provedor hospeda um aplicativo de modo centralizado e disponibiliza o acesso a vários clientes, pela Internet, em troca de uma taxa. Na prática, entretanto, as diferenças entre um aplicativo instalado no local do cliente e um aplicativo SaaS podem ser vistas em três dimensões: como aplicativo é licenciado, onde ele está localizado e como é gerenciado.

- **Licenciamento.** Em geral, os aplicativos instalados no local são licenciados para sempre, com pagamento único relativo a cada usuário ou local, ou (no caso de aplicativos construídos sob encomenda) de propriedade integral. Os aplicativos SaaS são licenciados, quase sempre, de acordo com um modelo de transação baseado no uso: cobra-se do cliente apenas as transações de serviço usadas. Existe também o modelo familiar da assinatura cuja base é o tempo: o cliente paga uma taxa fixa, por estação, por um determinado período, por exemplo, mensal ou trimestral, durante o qual terá direito ao uso ilimitado do serviço.
- **Localização.** Os aplicativos SaaS são instalados no local do provedor do SaaS, enquanto os aplicativos *on-premise*, naturalmente, são instalados no seu próprio ambiente de TI.
- **Gerenciamento.** Tradicionalmente, o departamento de TI é responsável por prestar serviços de TI aos usuários, ou seja, deve estar familiarizado com redes, servidores e plataformas de aplicativos, dar suporte e fazer diagnóstico de falhas e, ainda, resolver problemas de TI relativos à segurança, à confiabilidade, ao desempenho e à disponibilidade. Isso representa um grande volume de trabalho, e alguns departamentos de TI subcontratam algumas dessas responsabilidades de gerenciamento a terceiros, prestadores de serviços especializados em gerenciamento de TI. Na outra ponta do espectro, os aplicativos SaaS são completamente gerenciados pelo fornecedor ou pelo provedor de SaaS; na verdade, a implementação das tarefas e responsabilidades de gerenciamento não fica transparente para o consumidor. Os contratos de nível de serviço (SLAs) regem os compromissos de qualidade, disponibilidade e suporte a serem fornecidos pelo provedor ao assinante.

10.4. Considerações para Adotar o SaaS

Para qualquer aplicativo ou função determinada, pode-se determinar o estado de prontidão para SaaS, plotando as necessidades e as expectativas da empresa para cada item, usando a [Figura 10-1](#) como um guia.

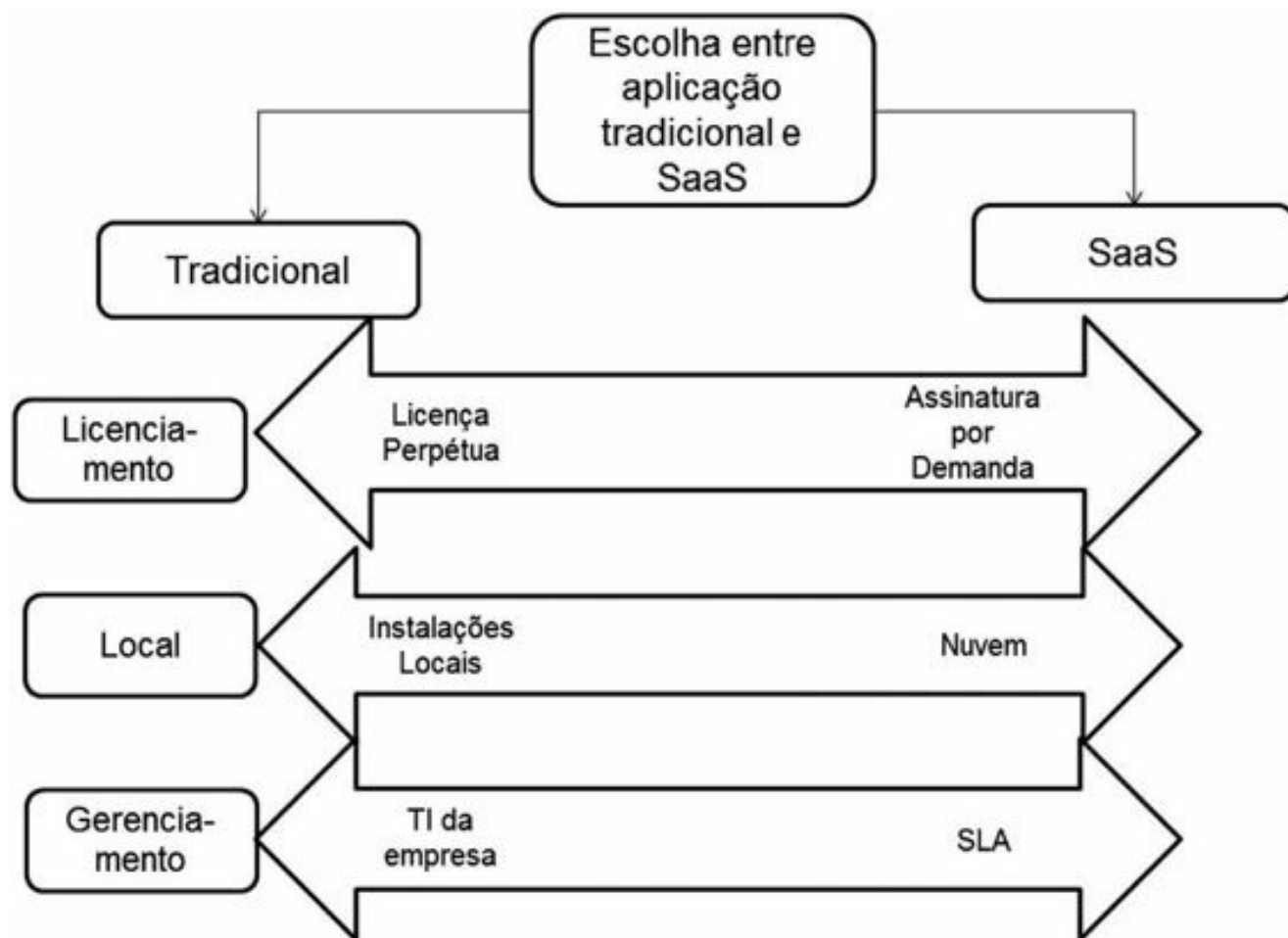


Figura 10-1 – Prontidão para SaaS

Se a empresa estiver próxima da extrema direita de cada item na [Figura 10-1](#), provavelmente a organização estará pronta para tentar mudar para SaaS. Se, ao contrário, estiver próxima da extrema esquerda, provavelmente terá de permanecer com a solução tradicional desse aplicativo, instalado no local. Qualquer outra combinação sugere que uma abordagem híbrida seria adequada; deve-se analisar o mercado para ver se é possível identificar soluções que sejam adequadas.

Encontrar a posição exata em cada sequência nas três dimensões envolve analisar várias considerações, as quais se resumem, ao final, no equilíbrio entre controle e custo. Algumas dessas considerações incluem:

- **Considerações políticas.** Às vezes, a decisão pode não ser possível devido a uma resistência interna da empresa, como, por exemplo, pessoas importantes que insistem que determinadas funcionalidades continuem internas, sob o controle do departamento de TI; as outras considerações, assim, tornam-se insignificantes. Implantações de *test-drive*, algumas vezes, ajudam a convencer gerentes avessos ao risco a aprovar projetos-piloto.

- **Considerações técnicas.** Os aplicativos SaaS, em geral, proporcionam alguma flexibilidade com relação à configuração do cliente, mas esta abordagem tem suas limitações. Se um aplicativo importante exigir conhecimento técnico especializado para sua operação e suporte, ou se exigir personalização que um fornecedor SaaS não possa oferecer, talvez não seja possível encontrar uma solução SaaS para esse aplicativo. Outro fator a ser considerado é o tipo e o volume de dados a serem transmitidos de/para o aplicativo, regularmente. A largura de banda da Internet fica muito aquém dos links Ethernet de 10 Gbits, normalmente encontrados nas LANs corporativas, e as transmissões de dados que levam poucos minutos na transferência entre servidores podem levar horas para transmitir de/para um aplicativo SaaS localizado no DATACENTER. Por isso, talvez seja mais sensato analisar uma solução que considere a latência da rede. Uma solução baseada em *appliances* (softwares e hardwares empacotados para uso no cliente), por exemplo, poderia implementar cache ou *batch*.
- **Considerações financeiras.** Considerar o custo total de propriedade (TCO) de um aplicativo SaaS, comparado com aquele de um aplicativo equivalente, instalado no local. Embora o custo inicial de aquisição dos recursos de software por SaaS seja normalmente inferior àquele dos aplicativos instalados no local, a estrutura de custo a longo prazo é mais incerta. Fatores que podem afetar o TCO de um aplicativo SaaS incluem o número de usuários licenciados, o volume de configurações personalizadas a ser executado para integrar o aplicativo SaaS em sua infraestrutura e, por fim, se os DATACENTERS existentes já produzem economia de escala, reduzindo, assim, a economia de custo em potencial do SaaS. Além disso, pode-se decidir retardar a implementação de um substituto SaaS daquele aplicativo caro ou recém-instalado, até que este produza um retorno sobre o investimento (ROI) satisfatório.
- **Considerações jurídicas.** Em várias partes do mundo, alguns setores estão sujeitos a regulação que impõe exigências quanto à emissão de relatórios e guarda de registros de dados que a solução SaaS a ser escolhida talvez não possa cumprir. É necessário considerar os ambientes regulatórios das várias jurisdições nas quais opera a empresa e como estes afetam as necessidades do seu aplicativo. Algumas vezes, as considerações técnicas e financeiras também têm ramificações legais, por exemplo, se os possíveis provedores de SaaS terão condições de atender os padrões internos de segurança e privacidade dos dados para evitar exposição legal.

10.5. Abordagens para a Arquitetura Multitenancy

10.5.1. Introdução

Pesquisas indicam que a falta de confiança é a principal razão para a não adoção do Software como Serviço (*Software as a Service* – SaaS). Os dados são bens importantes para qualquer negócio e no novo modelo estarão fora da organização, o que para muitas empresas é um problema.

A ideia que está por trás do SaaS é oferecer aos clientes um acesso centralizado às informações, por um custo menor, se comparado a uma aplicação executando localmente. Mas, desde

que se queira tirar vantagem de todos os benefícios do SaaS, uma organização precisa adaptar seus próprios dados a um determinado nível de controle e confiar no fornecedor de soluções de SaaS.

Para melhorar a confiança, uma das prioridades da arquitetura SaaS é a criação de uma arquitetura de dados que seja robusta e segura o bastante para satisfazer parceiros e clientes – preocupados com o controle dos dados empresariais (vitais para terceiros), ao mesmo tempo em que possuem uma boa relação de custo-benefício para gerenciamento e administração.

A arquitetura de dados é uma área na qual o nível de otimização de isolamento para uma aplicação SaaS pode variar significativamente – dependendo das considerações técnicas e de negócios. Arquitetos de dados mais experientes estão acostumados a considerar uma grande variedade de opções ao criar uma arquitetura para resolução de um caso específico, e o SaaS não é exceção, segundo Chong e outros (2006). Deve-se examinar três amplas abordagens, cada qual com diferentes posicionamentos sobre a extensão de um modelo isolado para um modelo compartilhado. O artigo “Arquitetura de Dados para Múltiplos Inquilinos”, dos autores citados (que atuam na Microsoft) é resumido aqui.

A **Figura 10-2** ilustra as três abordagens.



Figura 10-2 – Opções para Arquitetura de Dados

Cada uma das três abordagens apresenta vantagens e desvantagens.

10.5.2. Banco de Dados Separado

Armazenar dados de inquilinos em banco de dados separados é o modo mais simples de se obter um modelo de dados isolado.

Os recursos da máquina (e da aplicação) são geralmente compartilhados entre todos os inquilinos de um mesmo servidor, mas cada inquilino tem suas próprias informações, que permanecem isoladas – de maneira lógica – dos dados pertencentes a outros inquilinos. Os

metadados fazem a associação de cada um dos bancos de dados com o inquilino correto, e a segurança desse banco de dados evita que qualquer inquilino, acidentalmente ou não, acesse os dados de outro inquilino.

Atribuir um banco de dados para cada inquilino torna fácil a extensão do modelo de dados de uma aplicação para o atendimento às necessidades individuais desse próprio inquilino. Além disso, restaurar os dados de um inquilino, em caso de falhas, é um procedimento relativamente simples. Infelizmente, essa abordagem gera custos mais altos no gerenciamento do equipamento e backup dos dados. Os custos de hardware também são elevados – se comparados a abordagens alternativas – na medida em que houver um aumento no número de inquilinos hospedados nesse servidor – e este possuir alguma limitação no número de banco de dados suportados.

Separar os dados de cada inquilino em banco de dados individuais é uma abordagem recomendada se os requerimentos e custos relativos a hardware e gerenciamento o tornam apropriado aos clientes que estejam dispostos a pagar por mais segurança e customização.

10.5.3. Banco de Dados Compartilhado, Esquemas Separados

Outra abordagem envolve o armazenamento das informações de múltiplos inquilinos num mesmo banco de dados – cada inquilino com o seu próprio conjunto de tabelas, agrupadas em um esquema criado especificamente para o inquilino.

Quando um cliente inscreve-se pela primeira vez num serviço, o sistema de provisionamento cria um discreto conjunto de tabelas para o inquilino e o associa ao seu próprio esquema.

Da mesma forma que o método de isolamento, essa separação de esquemas é relativamente fácil de ser implementada, e os inquilinos podem, facilmente, estender o modelo de dados (tabelas são criadas de um conjunto de padrões, porém, uma vez criadas, elas não necessitam mais estar em conformidade com esse conjunto de padrões, e os inquilinos podem adicionar ou modificar colunas e, até mesmo, tabelas da forma que desejarem). Essa abordagem oferece um grau moderado da lógica de isolamento de dados e da segurança para os inquilinos, apesar de não ser igual ao modelo de isolamento total, e pode suportar um grande número de inquilinos por base de dados.

Um ponto fraco no esquema separado é que os dados do inquilino são mais difíceis de serem restaurados no caso de uma falha. Se cada inquilino possuir sua própria base de dados, restaurar uma única base de dados significa, simplesmente, restaurar o banco de dados através do backup mais recente. Com uma aplicação utilizando-se de um esquema separado, restaurar por completo um banco de dados pode significar a sobreposição dos dados de cada um dos inquilinos numa mesma base de dados com os dados do backup – sem levar em consideração se houve, ou não, perda. Além disso, para restaurar uma simples informação de um consumidor, o administrador do banco talvez tenha que restaurar a base de dados para um servidor temporário e depois importar as tabelas do cliente para um servidor de produção – uma tarefa potencialmente perigosa e demorada.

A abordagem através da separação de esquemas é apropriada para aplicações que usem um relativo (e pequeno) número de tabelas de banco de dados, na ordem de 100 tabelas por inquilino – ou menos. Essa abordagem pode acomodar mais inquilinos por servidor em relação a uma abordagem de banco de dados separados; assim, pode-se oferecer uma aplicação a um custo menor, enquanto os clientes aceitarem ter seus dados coalocados com os de outros clientes.

10.5.4. Banco de Dados Compartilhado, Esquemas Compartilhados

Uma terceira abordagem envolve o uso de um mesmo banco de dados e conjunto de tabelas para a hospedagem de múltiplas informações, de múltiplos inquilinos. Uma determinada tabela pode incluir registros vindos de múltiplos inquilinos, armazenados em qualquer ordem; e uma coluna com o ID de cada inquilino associa cada registro com o inquilino apropriado.

Das três abordagens apresentadas, a que se utiliza de um esquema compartilhado possui o menor custo de hardware e backup – por permitir servir o maior número de inquilinos por servidor. Entretanto, por causa desse compartilhamento, essa abordagem pode propiciar um esforço adicional de desenvolvimento na área de segurança – para garantir que inquilinos nunca possam se associar a dados de outros inquilinos, mesmo com a ocorrência de ataques ou bugs inesperados.

O procedimento para restauração de dados é similar à abordagem do esquema compartilhado, com uma complicação adicional: as linhas individuais dentro de um banco de dados de produção precisarão ser apagadas e então reinseridas no banco de dados temporário. Se houver um grande número de linhas nas tabelas afetadas, isso pode causar uma significativa queda de desempenho a todos os inquilinos que utilizam esse banco de dados.

A abordagem com esquema compartilhado é apropriada quando é importante que a aplicação seja capaz de servir um grande número de inquilinos com um pequeno número de servidores, e que possíveis clientes estejam dispostos a mudar de um modelo com isolamento de dados, em troca de menores custos – decorrentes da adoção desta abordagem.

10.5.5. Considerações Econômicas

Aplicações otimizadas por uma abordagem de compartilhamento requerem um grande esforço de desenvolvimento – em relação a aplicações criadas por uma abordagem de isolamento – (por causa da relativa complexidade do desenvolvimento em uma arquitetura compartilhada), resultando em custos iniciais maiores. Por causa de sua capacidade em suportar mais inquilinos por servidor, entretanto, os custos operacionais em uma abordagem de compartilhamento tendem a ser mais baixos.

Os esforços de desenvolvimento podem sofrer pressão de fatores econômicos e empresariais, influenciando na escolha da abordagem. O esquema com compartilhamento pode gerar uma grande economia financeira ao final de um grande projeto, mas requer um grande esforço inicial na fase de desenvolvimento, antes que esta possa gerar rendimentos. Se você não está disposto a empregar tanto capital na fase inicial de desenvolvimento, ou se precisa levar sua aplicação ao mercado o mais rápido possível, você pode considerar uma abordagem mais isolada.

10.5.6. Considerações de Segurança

Na medida em que a aplicação armazenar dados sigilosos dos inquilinos, estes terão grandes expectativas no que diz respeito à segurança, e o SLA necessitará prover garantias de segurança a esses dados. Uma concepção errada, e comum, baseia-se na ideia de que apenas um isolamento físico pode prover um nível de segurança apropriado. Dados armazenados através de uma abordagem com compartilhamento também podem contar com um forte esquema de segurança, mas requerem o uso de padrões mais sofisticados de projeto.

10.5.7. Considerações sobre Tenants

O número, a natureza e as necessidades dos tenants (inquilinos) que se espera servir afetarão a decisão sobre a concepção da arquitetura de dados de diferentes maneiras. As questões a seguir ajudam a definir o tipo de abordagem a ser escolhida:

- **Quantos inquilinos se espera ter?** As estimativas em relação à utilização da aplicação podem não estar claras. Mas pense em termos de magnitude: você está construindo uma aplicação para centenas de inquilinos? Milhares? Milhões? Mais? Quanto maior a expectativa em relação à base de inquilinos, maior a probabilidade de se considerar um modelo com compartilhamento.
- **Quanto espaço de armazenamento será ocupado, em média, pelos dados dos inquilinos?** Se você espera que algum dos inquilinos tenha uma enorme quantidade de dados, uma abordagem com uma separação dos dados talvez seja melhor (de fato, os requerimentos de armazenamento podem forçá-lo a adotar um modelo de banco de dados separado. Se for o caso, será muito mais fácil criar a aplicação de uma determinada maneira no começo e, posteriormente, mudá-la para a abordagem que separa o banco de dados).
- **Quantos usuários finais, em média, serão suportados?** Quanto maior o número, mais apropriado será o uso de uma abordagem com isolamento, para suprir a necessidade dos inquilinos.
- **Pretende-se oferecer algum serviço agregado por inquilino, como backup e restore?** Tais serviços são fáceis de serem oferecidos através de uma abordagem com isolamento.

A **Figura 10-3** ilustra as considerações feitas.

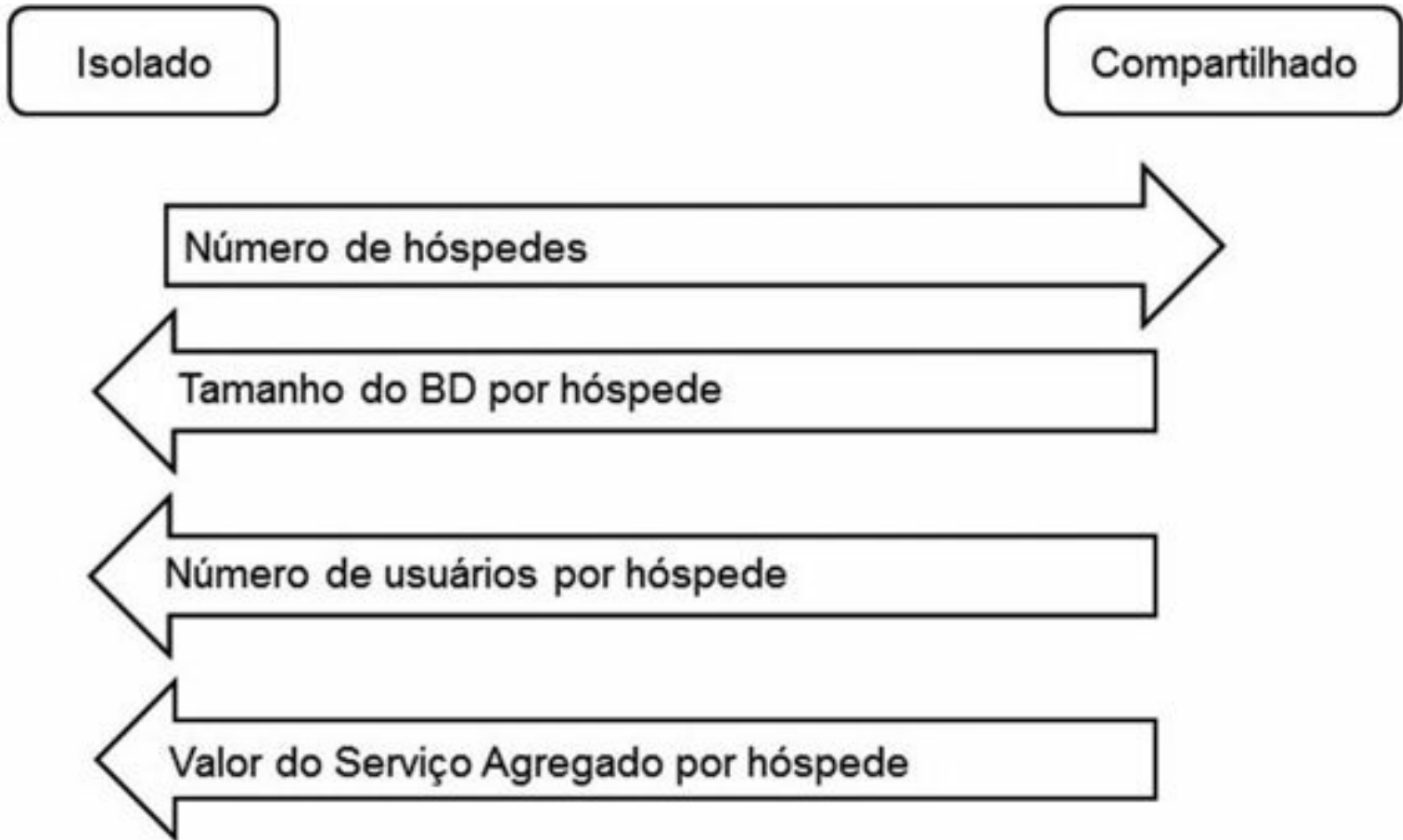


Figura 10-3 – Fatores relacionados aos Tenants

10.5.8. Considerações sobre Mudanças

Companhias, organizações e governos estão sujeitos a mudanças na lei que podem afetar as necessidades de segurança e armazenamento.

10.5.9. Considerações sobre Habilidades Necessárias

Criar uma única instância ou uma arquitetura para múltiplos inquilinos continua sendo uma habilidade relativamente nova, de modo que é difícil encontrar em profissional com experiência nesta área. Se os seus arquitetos e sua equipe de suporte não possuem um grande nível de experiência na construção de aplicações SaaS, será necessário adquirir o conhecimento para tal – ou você terá que contratar pessoas com experiência no assunto. Em alguns casos, uma abordagem com mais isolamento pode permitir à sua equipe um avanço nas habilidades de desenvolvimento tradicional de software em relação a uma abordagem com compartilhamento.

10.5.10. Qualidades de uma Aplicação SaaS

Os padrões apropriados para uma aplicação SaaS em relação à abordagem de banco de dados utilizada é mostrada na [Tabela 10-1](#). Em resumo, a abordagem para aumento da eficiência no ambiente compartilhado não pode comprometer a segurança.

Abordagem	Padrões de Segurança	Padrões de Extensibilidade	Padrões de Escalabilidade
Banco de dados separados	Conexões confiáveis a banco de dados Tabelas seguras de banco de dados Criptografia de dados do inquilino	Colunas customizadas	Aumento estrutural para um único inquilino
Banco de dados compartilhado, esquemas separados	Conexões confiáveis a banco de dados Tabelas seguras de banco de dados Criptografia dados do inquilino	Colunas customizadas	Particionamento horizontal baseado no inquilino
Banco de dados compartilhado, esquemas compartilhados	Conexões confiáveis a banco de dados Filtro de visualização do inquilino Criptografia de dados do inquilino	Campos pré-alocados Pares de nome-valores	

10.6. Conceito na Prática: [Force.com](https://www.force.com)

10.6.1. Introdução

Multitenancy é um *approach* que melhora a gerenciabilidade de aplicações SaaS. [Force.com](https://www.force.com) foi a primeira plataforma PaaS desenvolvida para criar aplicações SaaS *multitenancy*.

Aplicações dedicadas *single tenant* requerem um conjunto de recursos para satisfazer a uma organização.

Aplicações multitenant podem satisfazer as necessidades de múltiplos *tenants* (companhias, ou mesmo departamentos dentro das companhias) usando recursos de hardware e staff necessários para gerenciar uma instância única de software.

A [Figura 10-4](#) ilustra a ideia da aplicação *multitenant*.

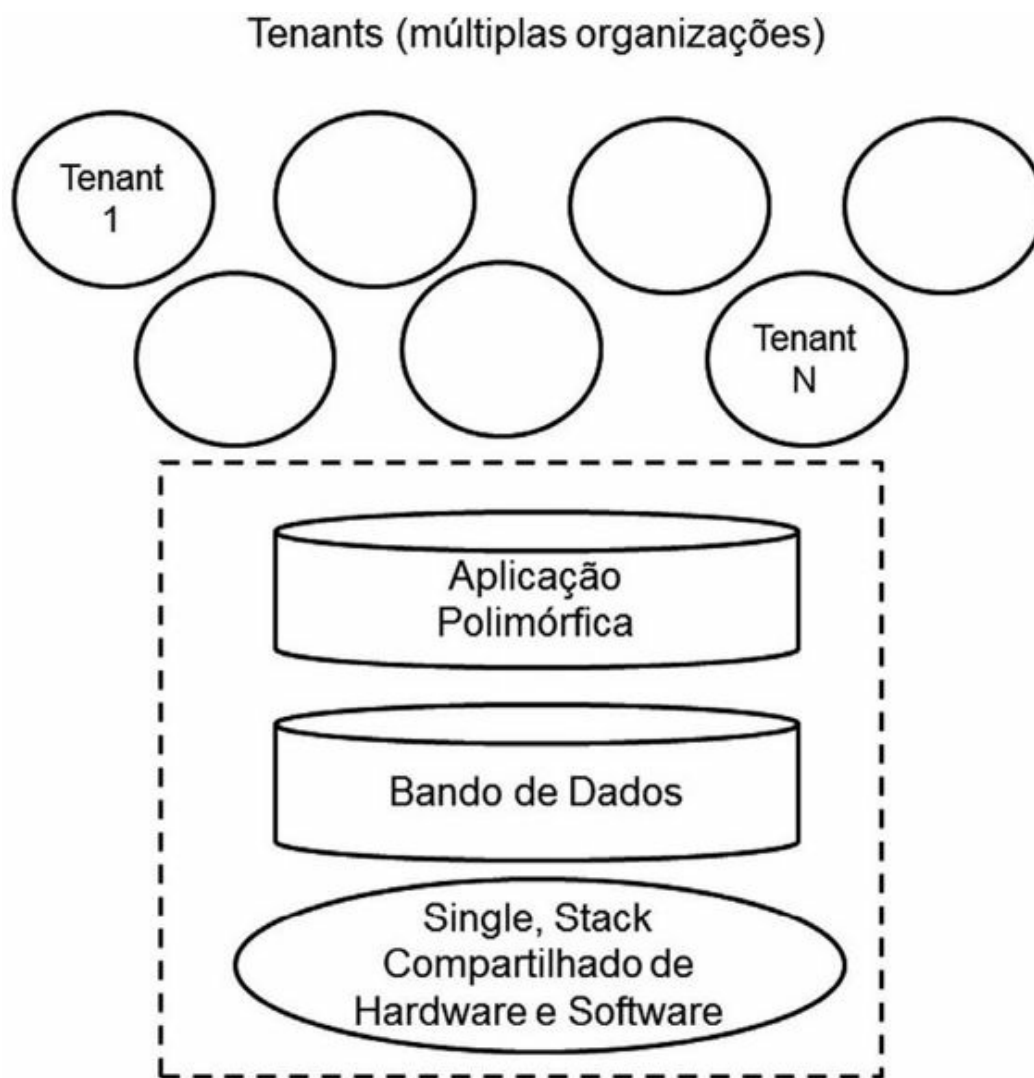


Figura 10-4 – Aplicação multitenant

Operar só uma instância da aplicação para múltiplas organizações dá uma grande economia de escala para o provedor. Somente um conjunto de hardware é necessário para atender as necessidades dos usuários. Um conjunto de hardware pode atender todos os usuários, e um staff experientado mais reduzido pode gerenciar todo o stack de software. Os desenvolvedores podem construir um único código em uma só plataforma.

Alguns fornecedores de SaaS evitam o desafio de desenvolver uma arquitetura *multitenant* e na verdade entregam solução de instâncias *single tenant* via IaaS.

10.6.2. Arquitetura Metadata-Driven

Multitenancy é prática quando permite suportar aplicações que são confiáveis, customizáveis, *upgradeable*, seguras e rápidas. Como uma aplicação *multitenant* possibilita que cada *tenant* crie extensões customizadas para objetos de dados padrões e novos objetos de dados customizados? Como dados específicos de um *tenant* podem ser mantidos seguros em um banco de dados compartilhado? Como pode um *tenant* customizar a interface da aplicação e a lógica do negócio em tempo real sem afetar a funcionalidade ou a disponibilidade da aplicação de outro *tenant*? Como é o tempo de resposta da aplicação com milhares de *tenants* subscrevendo o serviço?

A aplicação *multitenant* é dinâmica por natureza ou polimórfica para atender os vários *tenants*

e seus usuários.

A solução encontrada pela Salesforce foi desenvolver a arquitetura [Force.com](#). [Force.com](#) utiliza um *engine* de *runtime* que gera componentes de aplicação de metadados. Na arquitetura *metadata-driven* ilustrada na [Figura 10-5](#) existe uma clara separação entre *engine* de *runtime* compilado (kernel), dados de aplicação e metadados que descrevem a funcionalidade básica da aplicação e os metadados que correspondem aos dados e customizações de cada *tenant*. Estes contornos fazem os updates do kernel, modificações das aplicações ou a customização de componentes específicos acontecerem sem risco para os outros tenants.

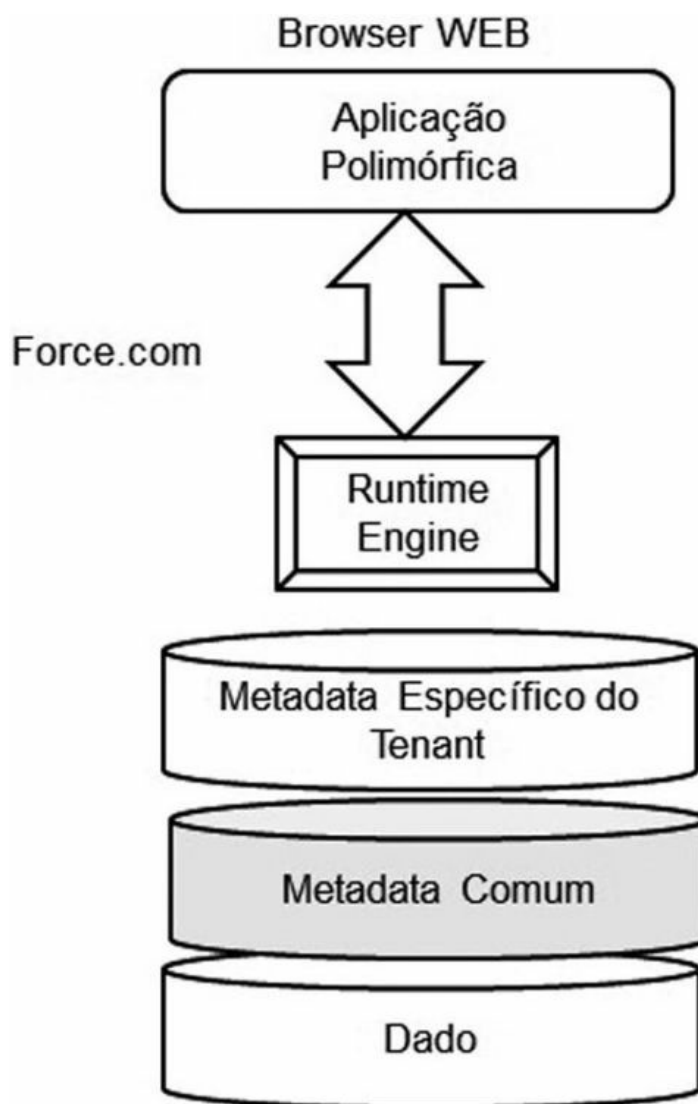


Figura 10-5 – Arquitetura [Force.com](#) MetaData-Driven

[Force.com](#) é uma nova arquitetura para desenvolvimento de aplicações na Internet. Frameworks anteriores não dão conta das novas demandas de aplicações *multitenants*.

10.7. Conceito na Prática: Office 365 da Microsoft

O pacote do Office continua a evoluir com o fornecimento de ferramentas mais abrangentes para ajudar usuários e maximizar a produtividade corporativa.

A proposta do Business Productivity Online Standard (BPOS) Suite foi a de ser um conjunto de ferramentas de colaboração e mensagens, fornecido como um serviço por assinatura, que oferece recursos avançados sem a necessidade de implantação e manutenção de software e hardware no local. Esses serviços online foram projetados para ajudar a atender necessidades de segurança, confiabilidade e produtividade do usuário.

O pacote do Office 365 é a próxima etapa nessa evolução. Diferentemente do BPOS, o Office 365 já incorporou o conceito de aplicação *multitenant*, portanto, foi construído para ser acessado por diversas organizações e seus diversos usuários. Em contraste, as aplicações BPOS são todas baseadas em versões *single tenant* do Exchange e Share-Point e não têm aplicações de produtividade do Office como Word, Excel e PowerPoint.

10.7.1. SharePoint Online

O Office 365 fornece um pacote de aplicativos baseados em nuvem para complementar e aprimorar a experiência do cliente da área de trabalho do Office.

O primeiro desses aplicativos é o SharePoint Online. Em vez de implantar e gerenciar o SharePoint internamente, o SharePoint Online fica hospedado fora dos limites empresariais. Assim, usuários empresariais têm acesso de qualquer lugar às funcionalidades de colaboração, gerenciamento de informações e *business intelligence* do SharePoint.

O SharePoint Online dá acesso a sites para:

- Gerenciamento e compartilhamento de arquivos e documentos importantes (My Sites).
- Manter as equipes em sincronia e gerenciar projetos importantes (Team Sites).
- Manter-se atualizado com as informações e as notícias da empresa (Intranet Sites).
- Compartilhar documentos de maneira segura com clientes e parceiros (Extranet Sites).
- Promover e realizar o marketing da empresa com um site voltado para o público.

É possível acessar o SharePoint Online diretamente em aplicativos cliente do Office 2010, inclusive o Word, o Excel e o PowerPoint. É possível criar documentos da maneira como você sempre fez e salvá-los em espaços de trabalho online do SharePoint.

De lá, pode-se acessar documentos por meio do SharePoint Online via web e também compartilhá-los com outros usuários, dentro ou fora de organização, por meio de controles de permissões granulares. Com o portal da web do SharePoint Online e o Office Web Apps, a coautoria e a colaboração também são possíveis em documentos em tempo real. Isso elimina a necessidade de revisões constantes por e-mail de um lado para outro.

O Office Web Apps – a versão online do Office – não necessita de Office no desktop para trabalhar. No entanto, esse cenário também significa que os usuários finais não podem se mover e sincronizar arquivos entre seu software Office e o Office Web Apps na nuvem.

Com o uso do My Sites do SharePoint, o SharePoint Online permite que os usuários adicionem áreas de interesse, habilidades e qualificações a seus perfis online, o que facilita encontrar especialistas no assunto dentro da organização. Isso não apenas melhora a produtividade corporativa, mas também a expansão de redes sociais e profissionais dentro da empresa. O SharePoint Online também dá suporte à marcação social sofisticada, garantindo que os usuários e as comunidades organizem as informações de maneira que façam mais sentido para seus negócios.

10.7.2. Exchange Online

O Office 365 também fornece infraestrutura de comunicação hospedada com o uso do Exchange Online e do Lync Online. O Exchange Online fornece infraestrutura de e-mail, calendário, contatos e tarefas, bem como toda a funcionalidade de uma implantação interna do Exchange, inclusive informações de presença, publicação de dados de disponibilidade e dicas de e-mail.

É possível acessar os serviços do Exchange Online de várias maneiras, inclusive o cliente de desktop do Outlook, o Outlook Web App e dispositivos móveis, com o uso do Exchange ActiveSync. O Outlook Web App fornece funcionalidade total entre navegadores, inclusive o Internet Explorer, o Safari e o Firefox.

Com referência a dispositivos móveis, o Office 365 dá suporte a Windows Phone, dispositivos Nokia, dispositivos Palm, iPhone e iPad da Apple e determinados telefones Android. O Exchange Online pode dar suporte a dispositivos BlackBerry em determinados planos. Esse nível de compatibilidade entre plataformas garante que você tenha acesso a informações críticas em qualquer lugar e a qualquer hora.

O Lync Online permite acessar uma versão hospedada do Microsoft Lync Server, o sucessor do Microsoft Office Communications Server. Ele fornece uma plataforma de comunicação em tempo real confiável e segura, além de uma grande quantidade de funcionalidades de próxima geração.

O Lync Online fornece comunicação em qualquer lugar e a qualquer hora por meio de mensagens instantâneas, chamadas de áudio e vídeo, reuniões online e webconferências. Você pode se mover diretamente entre conversas de mensagens instantâneas e reuniões online *ad hoc*, conduzir apresentações online para clientes ou parceiros com o uso de áudio, vídeo e compartilhamento de tela, convidar contatos externos para participar de reuniões online e exibir o status de presença de outros usuários no Lync e em outros aplicativos do Office.

O Exchange Online e o Lync Online dão suporte a um modelo de implantação de coexistência. Isso facilita a integração de serviços internos e hospedados. A coexistência permite flexibilidade para migrar para serviços baseados em nuvem em seu próprio ritmo. Você pode manter alguns usuários em servidores hospedados internamente e, ao mesmo tempo, migrar outros usuários para o Office 365. Mais informações sobre a arquitetura da implantação de coexistência podem ser obtidas na página Assistente de implantação do Exchange na Biblioteca TechNet.

10.7.3. Gerenciamento e Migração do BPOS para Office 365

Com o Office 365, paga-se pelo uso. Há três camadas de assinaturas disponíveis: Office 365 para empresas, Office 365 para pequenas empresas e o Office 365 para educação. A ideia é que cada uma das camadas forneça o conjunto adequado de serviços com base no tamanho e nos requisitos da organização. É possível provisionar facilmente novos usuários, à medida que eles entram na organização. Também é possível gerenciar usuários existentes com um Painel de Gerenciamento de TI online.

Independentemente de qual versão do Office 365 é escolhida, alguns requisitos básicos incluem atualização do desktop e do browser. Além disso, PCs Windows devem ter pelo menos o Windows XP SP3, Windows Vista SP2 ou Windows 7. Para a interação com o desktop baseado em nuvem Lync Online, o Office 365 também exige o novo cliente de desktop Lync 2010, a próxima versão do Office Communicator 2007 R2. Lync, que oferece mensagens instantâneas e reuniões on-

line, incluindo áudio e videoconferência.

Clientes com um Active Directory local que desejam sincronizar com seu diretório Office 365 também devem se certificar de que eles têm as versões necessárias de uma variedade de *server-side* software para produtos como Windows Server, .Net Framework e Windows Installer.

Os administradores também devem estar cientes de que, uma vez que a Microsoft inicie a migração de contas do BPOS para Office 365, o processo provavelmente leva 48 horas, durante o qual os administradores são bloqueados para fora da console de gerenciamento e itens SharePoint Online podem ser visualizados, mas não editados.

Além disso, as configurações BPOS para as políticas de retenção de caixa de correio, Exchange ActiveSync e Outlook Web App precisam ser reconfigurados no Office 365. Há também requisitos de software específico para computadores Linux e Mac OS, bem como para navegadores como Firefox, Chrome e Safari.

O Painel de Gerenciamento de TI permite adicionar novos usuários, redefinir senhas, ajustar camadas de assinaturas, criar grupos de segurança, gerenciar solicitações de serviços e assim por diante. Você também pode delegar controle administrativo em um nível granular. Isso permite dar a seus colegas acesso a funções administrativas específicas com base em funções de usuário. Por exemplo, pode-se conceder ao departamento jurídico acesso administrativo às caixas de entrada dos usuários.

O Office 365 também inclui licenciamento pré-pago e gerenciamento do Office Professional Plus. Isso caracteriza a última versão de aplicativos da área de trabalho do Office com o Office Web Apps integrado. Você e seus usuários podem acessar documentos, e-mail, calendários ou quaisquer outros documentos do Office virtualmente, em qualquer dispositivo.

O Office 365 permite alterar a maneira como os usuários criam, gerenciam e compartilham informações comerciais críticas. Por levar a experiência familiar do Office para a nuvem, fornecer acesso entre plataformas e dispositivos e por fornecer licenciamento flexível e funcionalidade de implantação, o Office 365 pode ajudar a empresa a se manter atualizada com um ambiente cada vez mais global e móvel.

10.8. Referências Bibliográficas

Architecting for the Cloud: Best Practices, January 2010. Last update January 2011.

Bridging the divide between SaaS and Enterprise Datacenters, Oracle, February 2010.

Frederick Chong, Gianpaolo Carraro. Software como Serviço (SaaS): Uma perspectiva corporativa. Microsoft Corporation, 2007. Carraro e Roger Wolter. Arquitetura de dados para múltiplos inquilinos, Microsoft Corporation, 2006.

<http://technet.microsoft.com/pt-br/magazine/gg558110.aspx>

SaaS Data Architecture. An Oracle White Paper, October 2008.

The [Force.com](#) Multitenant Architecture: Understanding the design of [Salesforce.com](#)'s Internet Application Development Platform, wp, [salesforce.com](#), 2008.

Frederick Chong, Gianpaolo Carraro. Software como Serviço (SaaS): Uma perspectiva corporativa,

2007.

- ¹ The big switch: rewiring the world, from Edison to Google. W. W. Norton Co, 2008.
- ² Ler o capítulo A Vantagem Essencial, do livro O Futuro da Administração, de Gary Hamel e Bill Breen, publicado em 2008 no Brasil pela Elsevier.
- ³ White, Tom. Hadoop: The Definitive Guide. 2009. 1st Edition. O'Reilly Media. Pág. 3.
- ⁴ Ver Reimaging IT: The 2011 CIO Agenda, Gartner.
- ⁵ Ver The Oracle Enterprise Architecture Framework, Oracle, 2009.
- ⁶ Seeding the Cloud: Enterprises Set Their Strategies for CLOUD COMPUTING, Forbes Insights, 2010.
- ⁷ IDCs Datacenter and Cloud Computing Survey, January 2010.
- ⁸ Ver livro DATACENTER: Componente Central da Infraestrutura de TI, página 62.
- ⁹ Ver Gerenciamento de Serviços de TI: Realidades e Tendência do uso do ITIL nas empresas brasileiras. Pesquisa FGV, 2007.
- ¹⁰ Ver texto DATACENTER em <http://www.furukawa.com.br/>¹¹ Ver The Multi-tiered Hybrid Data Center, HP, 2009.
- ¹² Modular/Container DATA CENTERS Procurement Guide: Optimizing for Energy Efficiency and Quick Deployment.
- ¹³ IDC Worldwide Expense to Power and Cool the Server Installed Base, 1996-2010
- ¹⁴ X86 server forecast by Ethernet connection type no IEEE 802.3 HSSG, April 2007, Interim Meeting.